An Exploration of the Sequence of a 2.9-Mb Region of the Genome of *Drosophila melanogaster*: The *Adh* Region

M. Ashburner,*,† S. Misra,‡ J. Roote,* S. E. Lewis,‡ R. Blazej,§ T. Davis, C. Doyle,§ R. Galle,§ R. George,§ N. Harris,§ G. Hartzell,‡ D. Harvey,‡,§§ L. Hong,‡ K. Houston,§ R. Hoskins,§ G. Johnson,* C. Martin,§,1 A. Moshrefi,§ M. Palazzolo,§,2 M. G. Reese,‡ A. Spradling,†† G. Tsang,‡,§§ K. Wan,§ K. Whitelaw,§ B. Kimmel,§,2 S. Celniker§ and G. M. Rubin§,‡,§§**

**Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, England, †EMBL—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, England, **Department of Pathology, University of Wales College of Medicine, Cardiff, CF4 4XN, Wales, ‡Berkeley Drosophila Genome Project, Department of Molecular and Cell Biology, University of California, Berkeley, California 94720-3200, §§Howard Hughes Medical Institute, Life Sciences Annex, University of California, Berkeley, California 94720, ††Howard Hughes Medical Institute, Carnegie Institution of Washington, Baltimore, Maryland and §Berkeley Drosophila Genome Project, Lawrence Berkeley National Laboratory, Berkeley, California 94720*

Manuscript received March 24, 1999
Accepted for publication June 15, 1999

ABSTRACT

A contiguous sequence of nearly 3 Mb from the genome of *Drosophila melanogaster* has been sequenced from a series of overlapping P1 and BAC clones. This region covers 69 chromosome polytene bands on chromosome arm *2L*, including the genetically well-characterized "*Adh* region." A computational analysis of the sequence predicts 218 protein-coding genes, 11 tRNAs, and 17 transposable element sequences. At least 38 of the protein-coding genes are arranged in clusters of from 2 to 6 closely related genes, suggesting extensive tandem duplication. The gene density is one protein-coding gene every 13 kb; the transposable element density is one element every 171 kb. Of 73 genes in this region identified by genetic analysis, 49 have been located on the sequence; *P*-element insertions have been mapped to 43 genes. Ninety-five (44%) of the known and predicted genes match a Drosophila EST, and 144 (66%) have clear similarities to proteins in other organisms. Genes known to have mutant phenotypes are more likely to be represented in cDNA libraries, and far more likely to have products similar to proteins of other organisms, than are genes with no known mutant phenotype. Over 650 chromosome aberration breakpoints map to this chromosome region, and their nonrandom distribution on the genetic map reflects variation in gene spacing on the DNA. This is the first large-scale analysis of the genome of *D. melanogaster* at the sequence level. In addition to the direct results obtained, this analysis has allowed us to develop and test methods that will be needed to interpret the complete sequence of the genome of this species.

> *Before beginning a Hunt, it is wise to ask someone what you are looking for before you begin looking for it.*
> Milne 1926

IT is nearly 100 years since W. E. Castle and his colleagues at Harvard University introduced *Drosophila melanogaster* to the joys and rigors of scientific research (Kohler 1994). From that slender beginning research with this small fly has dominated genetics and much of biology. It is, therefore, wholly appropriate that *Drosophila melanogaster* should join the new elite of organisms—as one whose genome will be sequenced in its entirety (Miklos and Rubin 1996; Rubin 1998). That goal is still some time away, but significant progress has already been made, with the determination of the complete sequences of the 338-kb *bithorax* and 430-kb Antennapedia regions (Lewis *et al.* 1995; Martin *et al.* 1995; S. Celniker, B. Pfeiffer, J. Knafels, C. Mayeda, C. Martin and M. Palazzolo, unpublished results) and with the availability of over 40 Mb of genomic sequence available in the public domain (Berkeley *Drosophila* Genome Project 1999; European *Drosophila* Genome Project 1999). There are many reasons, both pragmatic and theoretical, for wanting to complete the sequence of a model organism such as Drosophila. On a practical level, the availability of this sequence will be of immediate benefit to all studying particular genes. More theoretically, only by the completion of this sequence can we contemplate a description of the protein universe of Drosophila, can we answer with assurance the question of gene number in Drosophila, can we know the nature, number, and distribution of noncod-

This article is dedicated to the birth of Aden Misra Siebel, who waited so patiently to join us.

Corresponding author: Michael Ashburner, Department of Genetics, Downing St., Cambridge, CB2 3EH, England.
E-mail: m.ashburner@gen.cam.ac.uk

[1] *Present address:* JGI Sequencing Centre, Walnut Creek, CA 94598.

[2] *Present address:* Amgen Inc., Thousand Oaks, CA 91320.

ing regions of DNA (including transposable elements), or can we explore the Drosophila genome for regularities in sequence organization that may correlate with chromosome organization. Moreover, the availability of the complete sequence of Drosophila will itself be a major impetus to evolutionary studies and to comparative insect genomics. Finally, but by no means least important, the sequence itself will spur functional studies, themselves of great interest to all biologists, especially those struggling to interpret the function of genes of the larger genomes of mammals.

The analysis and interpretation of long genomic sequences pose several unsolved problems, among which are gene prediction and correlation of genetically identified loci with computationally predicted genes. We have selected the 2.9-Mb *Adh* region, a region of the genome of *D. melanogaster* that was already well characterized by conventional genetic analyses, as a test-bed to develop and evaluate approaches to large-scale genomic sequence annotation in Drosophila. This chromosome region is defined as the 69 polytene chromosome bands from 34C4 to 36A2 on chromosome arm *2L*, which is the region between (and including) the previously known genes *kuzbanian* (*kuz*) and *dachshund* (*dac*). Genetic analysis of this chromosome region began with the studies of E. H. Grell in the early 1960s and the recovery of an *Adh*$^-$ deletion, *Df(2L)64j* (Grell *et al.* 1968). W. Sofer and students, especially J. M. O'Donnell (O'Donnell *et al.* 1977), recovered several more deletions, using formaldehyde as a mutagen, and defined 12 loci by complementation analysis among 33 EMS-induced lethal mutations uncovered by these deletions. These studies have been continued in the last 20 years by M. Ashburner's group (*e.g.*, Woodruff and Ashburner 1979a,b).

Genetic analysis has defined 73 genes in this chromosome region. Of these genes, 65 are represented by mutant alleles and 8 more are predicted on the basis of the phenotypes of overlapping deletions. Of those with mutant alleles, 50 genes have at least one lethal allele (*i.e.*, they are genes whose activities are vital), 6 are known only from sterile alleles (2 male sterile and 4 female sterile), 8 only from alleles with clear visible phenotypes, and 2 genes have alleles with no gross phenotype: *Adh* and *smi35A*. Forty-nine protein-coding genes (and 5 tRNA genes) in this region had been molecularly characterized prior to or during our work; these included 7 that had not been identified by genetic analysis. In addition to a collection of over 1038 different mutant alleles of genes in this region, the genetic analysis was enormously aided by a very large collection of chromosome aberrations, including 86 inversions, 109 translocations, 317 deletions, and 40 duplications. Apart from some conventional recombination mapping in the early stages of the project, all genes have been ordered by deletion mapping. The genetic positions of the breakpoints of many inversions and translocations have been mapped with respect to the genes, often by combining these breakpoints with others to synthesize deletions or duplications.

These genetic data posed two major questions. The first was that of "saturation": What proportion of the genes had been identified by the genetic analysis? It is well known (*e.g.*, Barrett 1980) that the distribution of mutant hits to genes defies any rigorous statistical estimation of the size of the class of genes that are mutationally silent (see Lefevre and Watkins 1986). This is particularly true in the present case, since many independent mutagenesis screens using a variety of deletions have been done, as have several specific locus screens. These mutation screens have been done with a variety of chemical agents, with ionizing radiation and with *P* elements, and although the most mutable genes in general screens have 50 or more alleles (*e.g.*, *wb* and *esg*), we already know, or predict, some genes that have been refractory, including those eight genes predicted from overlapping deletion phenotypes. Moreover, we had no experimental estimate of the number of genes that give no phenotype when mutant (see below). The second question is that raised by the very nonrandom clustering of aberration breakpoints. There are two extreme interpretations of this clustering: that the different regions differ in target size or that there is some intrinsic property that biases the recovery of chromosomal breaks. Both this question, and that of "saturation," have been answered from the analysis of the sequence of this region.

There is direct experimental evidence, or prediction, for 229 genes in the 2.9 Mb of sequenced DNA. Of these, there is evidence for function or some hint of function from sequence matches for 102 genes. One of the challenges for the future is to discover, by experiment, the function of all of the genes.

MATERIALS AND METHODS

Genetics: All of the mutations and chromosome aberrations used in this study are fully described in FlyBase (FlyBase Consortium 1999). Table 1 presents a summary of the mutations that have been identified. The majority of these have been published in previous articles from M. Ashburner's laboratory, and others have been given to us by colleagues; those that are new are described in FlyBase. Where possible we have mapped aberration breakpoints genetically by combining the elements of translocations (by segregation) or inversion breakpoints (by recombination, using autosynaptic intermediates in the case of pericentric inversions; see Gubb 1998) so as to synthesize deletions whose limits could be mapped by complementation. All genetic crosses were, unless otherwise stated, done between balancer heterozygotes and care was always taken to allow any very delayed progeny to eclose. A failure of complementation is based upon the absence of nonbalancer progeny, usually in progenies of 200 flies or more. Crosses were routinely done on standard laboratory food at 25°.

P elements from several laboratories, from screens for lethal *P* elements on chromosome *2* (see Spradling *et al.* 1995),

TABLE 1

Genes in the *Adh* region identified by genetic analysis

			Mutant alleles				
Gene symbol	Class[a]	Sequence[b]	EMS	Radiation[c]	*P*	Aberration	All[d]
kuz	l	+	0	0	9	4	20
l(2)34Db	l	−	12	0	0	1	14
Sos (l(2)34Ea)	l	+	25	0	4	0	40
b	v	+	8	105	0	13	144
tamas (l(2)34Dc)	l	+	8	0	0	0	8
Sop2 (l(2)34Dd)	l	+	3	0	0	0	3
Orc5 (l(2)34Df)	l	+	2	0	0	0	2
MtPolB (l(2)34De)	l	+	3	0	0	0	3
RpII33 (l(2)34Dg)	l	−	2	0	1	0	3
l(2)34Ec	l	−	0	0	0	0	0
Ance (l(2)34Eb)	l	+	2	0	0	0	2
j	v	−	2	0	0	0	7
rk	v	+	5	3	1	1	19
l(2)34Fa	l	−	2	0	1	1	4
smi35A	nv	−	0	0	3	0	3
wb (l(2)34Fb)	l	−	32	2	11	3	52
ms(2)34Fe	ms	−	0	0	0	0	1
l(2)34Fc	l	−	11	0	0	0	11
l(2)34Fd	l	−	10	0	0	0	10
l(2)35Aa	l	+	7	0	0	0	7
elB	v	−	0	1	3	5	9
pu	v	−	6	0	0	0	7
elA	v	−	8	6	0	1	16
noc	l	+	9	5	2	5	22
osp	v	+	13	7	2	9	33
Adh	nv	+	18	10	0	0	106
ms(2)35Bi	ms	−	0	0	0	0	0
l(2)35Bb	l	−	8	1	1	0	11
l(2)35Bf	l	−	9	0	0	0	9
l(2)35Bc	l	−	8	0	1	0	9
l(2)35Be	l	−	7	0	0	0	7
l(2)35Bd	l	−	6	0	1	1	8
l(2)35Bg	l	−	2	0	1	0	4
Su(H) (l(2)35Bh)	l	+	18	0	2	0	30
ck	l	+	18	0	2	1	24
TfIIS (l(2)35Cf)	l	+	0	0	0	0	2
vas	fs	+	14	0	3	0	23
stc (l(2)35Cb)	l	+	5	0	2	2	9
rd	v	−	9	3	0	0	12
l(2)35Cc	l	−	0	0	0	0	0
gft (l(2)35Cd)	l	+	7	1	1	0	11
ms(2)35Ci	ms	−	0	0	1	0	1
esg (l(2)35Ce)	l	+	3	0	54	5	72
l(2)35Cg	l	−	0	0	0	0	0
worniu (l(2)35Da)	l	+	4	0	0	0	8
l(2)35Ch	l	−	0	0	0	0	0
sna (l(2)35Db)	l	+	14	6	0	2	24
lace (l(2)35Dc)	l	−	14	0	6	1	24
CyeE (l(2)35Dd)	l	+	3	0	9	1	16
l(2)35Df	l	−	5	0	1	0	6
l(2)35Di	l	−	1	0	0	0	1
Gli (l(2)35Dg)	l	+	5	0	9	0	19
l(2)35Ea	l	−	1	0	1	0	4
l(2)35De	l	−	0	0	0	1	1
l(2)35Dh	l	−	1	0	0	0	1
fs(2)35Ec	fs	−	0	0	0	0	0

(*continued*)

TABLE 1

(Continued)

			Mutant alleles				
Gene symbol	Class[a]	Sequence[b]	EMS	Radiation[c]	*P*	Aberration	All[d]
ms(2)35Eb	ms	−	0	0	0	0	0
fs(2)35Ed	fs	−	0	0	0	0	0
BicC	fs	+	15	0	0	2	17
beat	l	+	4	0	0	2	6
Ca-α1D (l(2)35Fa)	l	+	3	0	0	0	4
twe	l, ms, fs	+	0	0	2	2	5
crp (l(2)35Fd)	l	−	1	0	21	0	24
l(2)35Fb	l	−	1	0	0	0	2
heix (l(2)35Fc)	l	−	1	0	3	0	5
Sed5 (l(2)35Ff)	l	+	1	0	0	0	1
cni	fs	+	3	0	0	0	4
fzy	l	+	3	1	1	0	10
cact	l	+	26	2	3	4	59
l(2)35Fe	l	−	1	0	1	0	2
chif	fs	+	7	0	3	0	10
l(2)35Fg	l	−	0	0	1	0	1
dac (l(2)36Ae)	l	+	0	0	1	0	8

[a] The most "extreme" mutant phenotype known. v, visible; l, lethal; ms, male sterile; fs, female sterile; nv, none of these. Only ADH-null alleles of *Adh* are included in this table.

[b] + indicates that a genomic or cDNA sequence of the gene was determined independently of this work. Except for the following these are all available from the nucleic acid sequence databases: *l(2)35Aa* (C. Flores), *Sop2* (A. Hudson), *gft* (H. Mistry), *rk* (J. Baker), *ck* (D. Kiehart), *vig* (K. Edwards), *worniu* (T. Ip), and *chiffon* (G. Landis and J. Tower). Sequences of the following known genes that have not been identified genetically are also available from the sequence databases: *Mst35Ba, Mst35Bb, tRNA:G3:35Ba–35Be, spl1, ppk, Adhr, B4, Rab14, Tim17, PRL-1, Idgf1, Idgf2,* and *Idgf3*. In addition cDNA sequences of the following new genes have been determined by others: *BG:DS01514.2* and *BG:DS05899.1* (M. Leptin), *beat-B* and *beat-C* (T. Pipes).

[c] X rays, γ-rays, and neutrons.

[d] Includes alleles induced by other mutagens, *e.g.* other chemicals, UV light, alleles of unknown origin, and spontaneous alleles. Genes with no mutant alleles are those predicted on the basis of overlapping deletion phenotypes (see text). Aberration alleles include those in the various mutagen columns. *P*-element alleles do not include those induced by PM dysgenesis or transposase-induced derivatives of *P*-element alleles (but these are included in the totals).

were screened against three deletions that, in sum, cover the entire genetic interval of interest—(*Df(2L)b84a7*, *Df(2L)A48*, and *Df(2L)r10*)—and then mapped more precisely using appropriate deletions and mutant alleles. We are very grateful to I. Kiss for the preliminary screen with his *P*-element collection. Further *P* elements were initially identified only on the basis of the chromosomal mapping of their insertion site by *in situ* hybridization to polytene chromosomes, using a *P*-element probe and standard techniques. These were then subjected to genetic analysis, typically tests for complementation with appropriate deletions and mutant alleles representative of candidate loci. The EP lines used in this study were from the collection described by Rørth *et al.* (1998).

P-element excisions and male recombinants were generated using *P{Δ2-3}99B* as the source of an active *P* transposase. These derivatives were then characterized by conventional genetic complementation analyses.

Cytology: For conventional polytene chromosome analysis we used propionic-carmine-orcein squash preparations. *In situ* hybridization was performed by standard procedures using biotinylated probes and horseradish peroxidase staining. Polytene chromosomes were interpreted using the revised maps of C. B. and P. N. Bridges (see Lefevre 1976).

Clones: The P1 clone library, with an average insert size of 80 kb, was that prepared from an isogenic *y*; *cn bw sp* stock in the vectors pNS583tet14Ad10 and pAd10sacBII (Sternberg 1990) and described by Smoller *et al.* (1991). The strategy for building contigs of overlapping clones has been described by Kimmerly *et al.* (1996). The first stage was to build a "framework" map of the genome of *D. melanogaster* by mapping over 2600 of the P1 clones to the polytene chromosomes by *in situ* hybridization (Hartl *et al.* 1994). Then, short sequence tagged sites (STS) were used to determine overlaps between P1 clones by STS-content mapping, using a PCR-based approach (Olson *et al.* 1989; Green and Olson 1990). STS sequences were derived from a number of sources: end sequences of P1 clones, insertion sites of *P* elements determined after plasmid rescue or inverse PCR, and sequences of known Drosophila genes. BAC clones were from a newly constructed library in pBACe3.6 (Osoegawa *et al.* 1998; K. Osoegawa, A. Mammoser and P. de Jong, unpublished results). This is a 20-hit library from a partial *Eco*RI digestion of DNA from the *y*; *cn bw sp* isogenic stock.

The P1 clones were first assembled into eight contigs by screening a 5-hit P1 clone library. By generating STS sequences determined from the ends of these contigs, and then mapping these to a second larger P1 clone library (10 hit), and by directed PCR experiments, these seven contigs assembled into two, of 0.8 Mb and 1.9 Mb, plus an isolated P1 clone containing the *kuzbanian* gene. The gaps between the two long contigs and between the isolated P1 clone and the 1.9-Mb contig were closed by screening the BAC clone library with sequences prepared from the appropriate end clones.

DNA sequencing: The sequence of the *Adh* region has been assembled by first determining the sequences of the 51 individual P1 clones that comprise the 0.8-Mb and 1.9-Mb contigs. The gap between the two contigs was filled by sequencing the BAC clone BACR44L22. The gap between the P1 clones DS07660 and DS01368 was filled by sequencing BACR48E02. Table 2 lists the clones sequenced and their DDBJ/EMBL/GenBank accession numbers.

The sequencing strategies have evolved over time. Essentially, *ca.* 3-kb subclone libraries of randomly sheared DNA were prepared from each P1 clone in plasmid vectors. The sequences of both ends of each plasmid insert were determined using primers complementary to the vector and these sequences were used to assemble a set of overlapping 3-kb clones that span an entire P1 clone. The 3-kb clones were then sequenced using a combination of transposon-mediated sequencing (Kimmel *et al.* 1997) and custom oligonucleotide-primed sequence runs. All sequences were determined on both DNA strands and assembled using the PHRAP program (P. Green, unpublished results). The error rate was estimated using PHRAP quality scores as <1 in 10,000. We wrote our own genomic assembler to generate a single complete sequence of the entire region from the individual clone sequences. The core alignment software used in this assembler was the sim4 program of Florea *et al.* (1998). The assembler iteratively runs sim4 against pairs of sequences that are known to overlap from the physical mapping data. The assembler then uses the exact alignment that covers the two ends of the clones to incrementally construct the complete sequence, performing reverse complementation when needed.

cDNA identification and sequencing: cDNA clones derived from genes in the 34D-36A region were identified by searching for sequence matches between the genomic DNA sequence and 5′ expressed sequence tags (ESTs) from the Berkeley Drosophila Genome Project (BDGP)/Howard Hughes Medical Institute (HHMI) Drosophila EST project (http://www.fruitfly.org/EST/). In addition, cDNAs corresponding to *crp*, *heix*, *l(2)35Fe*, *anon-35Fa*, *anon-35F/36A*, *BG:DS02740.2*, *BG:DS02740.4*, *BG:DS02740.8*, *BG:DS02740.9*, and *BG:DS02740.10* were isolated by screening the LD cDNA library using the method of Munroe *et al.* (1995). The LD cDNA library was made from poly(A)$^+$-selected RNA from 0–22-hr embryos, size fractionated ($\sim$1 to 6 kb), and directionally cloned in either the Stratagene (La Jolla, CA) Uni-Zap XR vector or the pOT2 plasmid (both *Eco*RI/*Xho*I digested; L. Hong, unpublished results). For each gene, the longest available cDNA was sequenced from one strand to allow unambiguous alignment with the genomic sequence. The cDNA sequences were aligned with the genomic sequence using the sim4 program of Florea *et al.* (1998). Because these cDNA sequences were low-pass, single-stranded sequence it was not always possible to construct a single open reading frame from sim4 alignments. In those cases, adjustments were made by an annotator. The virtual cDNA sequences were verified using the ORFfinder program (v. 0.1, E. Frise, unpublished results) and their structures relative to the genomic sequence manually checked in CloneCurator (see below).

Molecular mapping of *P*-element insertion sites: The precise insertion sites of all *P* elements described here were determined by comparison of the reference genomic sequence with a sequence that spanned the junction between a *P* element and the genome using sim4. These junction sequences were determined from either plasmid-rescued clones or inverse PCR products, as described in Spradling *et al.* (1999). The insertion site is reported as the first base pair of the 8-bp target site duplication generated by the *P*-element insertion.

TABLE 2

Sequenced P1 and BAC clones in region 34D-36A

Clone	*In situ*	Accession no.
BACR48E02	—	AC006302
DS07660	—	AC003924
DS01368	34C4-34D2	AC002434
DS08249	34D1-34D2	L49405
DS08284	—	AC004348
DS00941	—	AC001659
DS08220	34D6-34E3	AC001664
DS00180	34E4-34E5	AC001660
DS01514	34F1-34F2	AC002515
DS00131	34E4-34E5	AC001662
DS05899	34F1-34F2	AC004326
DS01759	—	AC004360
DS01523	34F3-34F4	AC003120
DS01652	—	AC001666
DS03792	35A1-34A2	AC001661
DS01068	—	AC002516
DS06238	35B2-35B3	AC004118
DS08340	—	AC001663
DS04641	—	AC002440
DS01160	—	AC001665
DS01486	35B6-35B7	AC004359
DS09219	—	AC001647
DS07721	35B2-35B10	L49403
DS00810	35B6-35B7	L49404
DS06874	—	AC001657
DS03431	—	AC001648
DS03144	35B6-35B7	AC001649
DS03323	—	AC002439
DS01219	—	AC004244
DS00929	35B8-35C1	AC002502
DS04929	—	AC003696
DS03192	35C1-35C2	AC004545
DS09194	35C1-35C3	AC004545
DS07295	35C1-35C2	AC004545
DS05639	—	AC002437
DS07851	—	AC004361
DS01362	35D1-35D2	AC002436
DS03023	35D1-35D2	L49394
DS01845	35D1-35D3	AC001646
DS04862	35D3-35D4	AC003698
BACR44L22	—	AC006303
DS07108	—	AC004362
DS09217	—	AC003700
DS02252	35E3-35E5	AC002493
DS00365	35F1-35F4	AC004113
DS07486	35F1-35F2	AC003925
DS08681	35F1-35F2	AC001651
DS00913	35F1-35F2	AC001658
DS04095	—	AC002501
DS02795	—	AC002441
DS07473	—	AC003701
DS02740	—	L49408
DS09218	35F11-36A2	AC002438
DS02780	36A1-36A2	AC002514

Clones are listed in their physical order along the chromosome from distal to proximal. An *in situ* hybridization site to polytene chromosomes is only given if this has been determined directly (rather than being inferred from the contig).

Sequence analysis: Two broad categories of computational method were used together to predict and identify genes. The first was gene prediction algorithms, based on the statistical properties of protein-coding regions. The second category of method used alignment algorithms for predictions based upon similarities of the sequence with other sequences in the public domain, both nucleic acid and protein.

The main gene prediction program used in the early stages of this analysis was GENEFINDER (v. 0.83; Green 1995), trained on a Drosophila sequence data set (G. Helt, unpublished results). GENEFINDER predicts genes on the basis of the statistical properties of their sequence, codon usage, codon preference, and splice site profiles. More recently, we made a comparison of the performance of a number of different programs using the sequence of the P1 clone DS02740. This showed that GENSCAN (v. 1.0; Burge 1997; Burge and Karlin 1997), trained on a vertebrate sequence data set, gave more reliable predictions than GENEFINDER, GENIE (Reese *et al.* 1997), or a version of GRAIL trained on a Drosophila sequence training set (Xu *et al.* 1995). This comparison showed a tendency for GENSCAN to overpredict genes. This characteristic was complemented by GENEFINDER, which tends to underpredict genes. For this reason, both programs were used for the final data analyses, using their default parameters. Predictions with scores lower than 45 for GENSCAN or 20 for GENEFINDER were ignored. No current gene prediction program behaves well with introns that are either very large or very small, and these errors were corrected, whenever possible, by using available alignment data. tRNA genes were predicted using the tRNAscan-SE program (v. 1.02) of Lowe and Eddy (1997).

To estimate the statistical properties of *D. melanogaster* protein-coding regions a nonredundant data set of coding regions (CDS) was made. By nonredundant we mean that for any one gene only one CDS is included, even if the gene encodes multiple protein products (that included was usually the longest complete sequence available from the EMBL Nucleic Acid Sequence Data Library). All of the CDS regions were checked for legitimate start and stop codons and for a continuous open reading frame in between these. Four genes with non-ATG starts were included in this data set (CTG, *amn*, *ewg*; GTG, *Cha*; CTC, *cpo*) following advice from D. Cavener, as were two CDSs (*oaf* and *kelch*) with in-frame UGA codons, perhaps coding for seleno-cysteine. This data set of 1335 CDSs was used for the construction of normalized codon and di-codon (hexamer) tables (Helt 1997) and is available as cds_sequence_set.embl.v1.5 from ftp://ftp.ebi.ac.uk/pub/databases/edgp/sequence_sets/ and as na_embl.dros.v1.5 from http://www.fruitfly.org/sequence/download.html.

Databases against which similarity searches were made included GenBank, dbEST, SWISS-PROT, SPTREMBL, and sequences from the European *Drosophila* Genome Project (EDGP). Updates of these were collected weekly, the sequence data sorted into species-specific files, and all submissions from the Berkeley *Drosophila* Genome Project removed to provide data sets for searches. These data sets were then processed to append all database cross-references to FASTA header lines. For sequence similarity searches the BLASTN, BLASTX, and TBLASTX programs (version 2.0a) of W. Gish (unpublished results) were used (with the option $B = 1{,}000{,}000$, options filter = SEG + XNU).

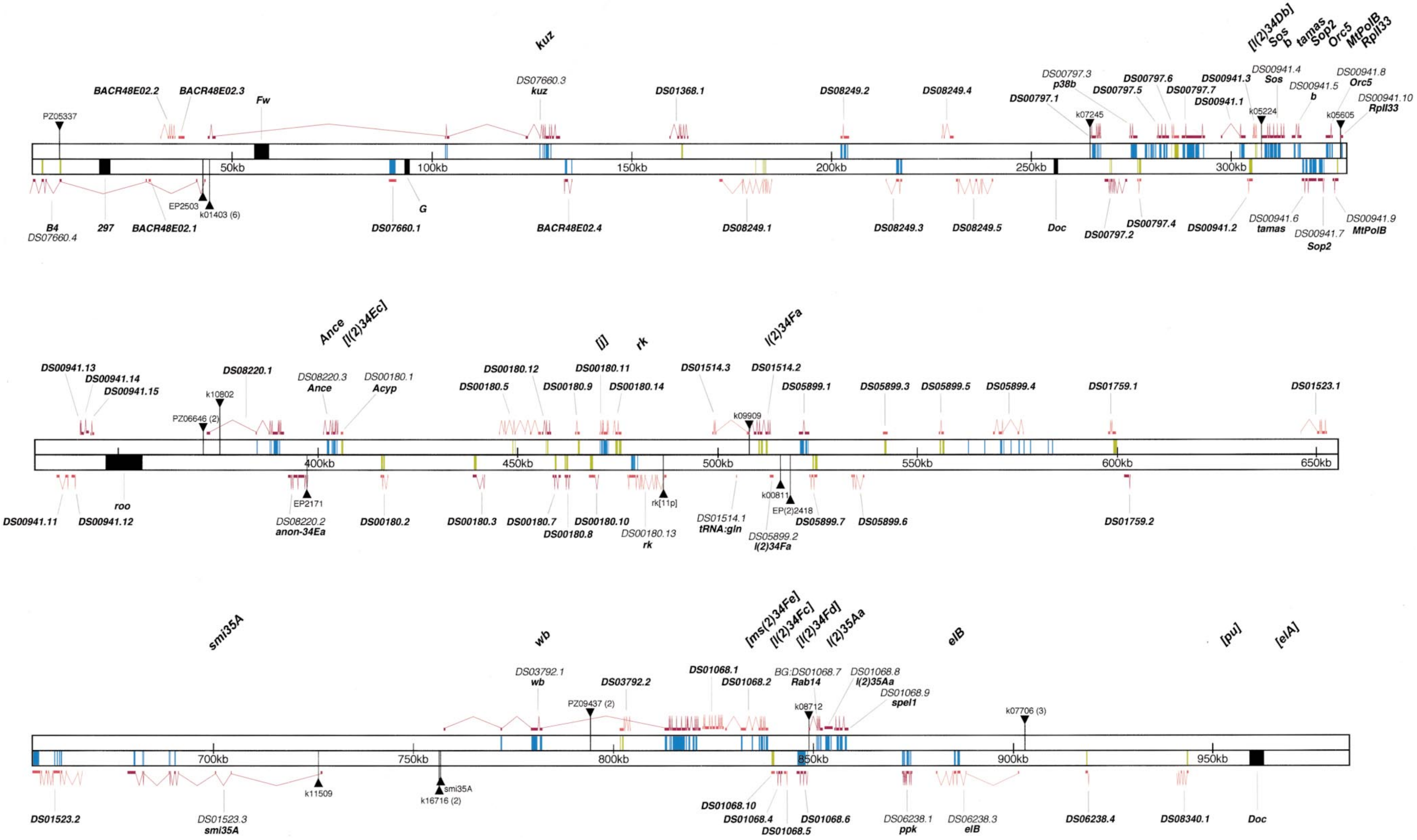
Transposable elements were screened using a nonredundant data set of transposable element sequences from which all "flanking" DNA sequences had been trimmed. This data set was originally derived from the EMBL Nucleotide Sequence Data Library records, but as our analysis progressed more complete sequences of elements only known before from partial sequence were added, replacing incomplete sequences. This data set is available from ftp://ftp.ebi.ac.uk/pub/databases/edgp/sequence_sets/transposon_sequence_set.embl and from http://www.fruitfly.org/sequence/download.html (as na_te.dros).

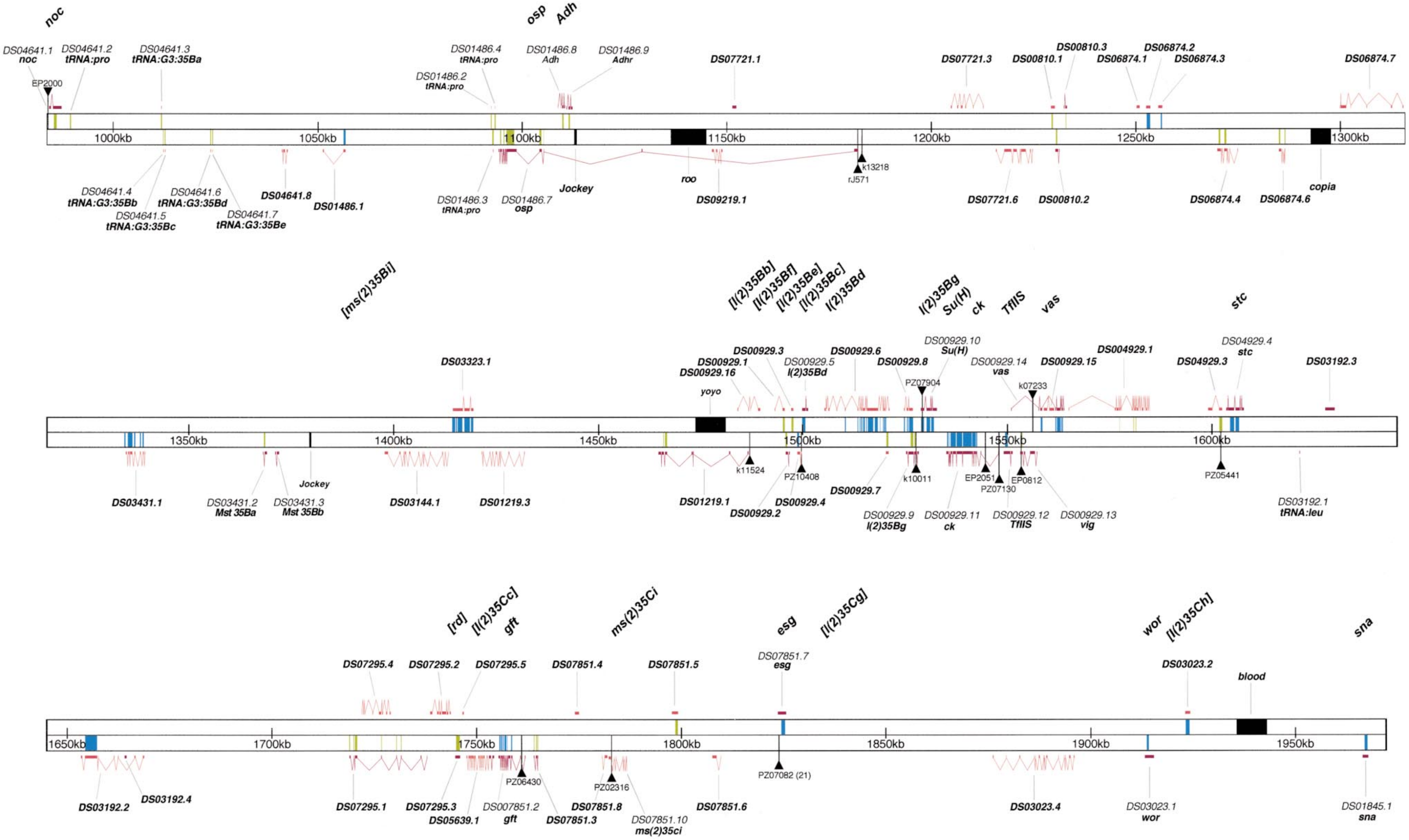
A collection of repetitive sequences from *D. melanogaster*, not otherwise included in the transposable element sequence set, was also made. This data set includes, *e.g.*, satellite DNA sequences and a miscellany of sequences annotated as being repetitive by FlyBase. It is not as nonredundant as the other two data sets, and was only used for screening for sequences similar to those previously described as repetitive. The data set is available from ftp://ftp.ebi.ac.uk/pub/databases/edgp/sequence_sets/repeat_sequence_set.embl and http://www.fruitfly.org/sequence/download.html (as na_re.dros).

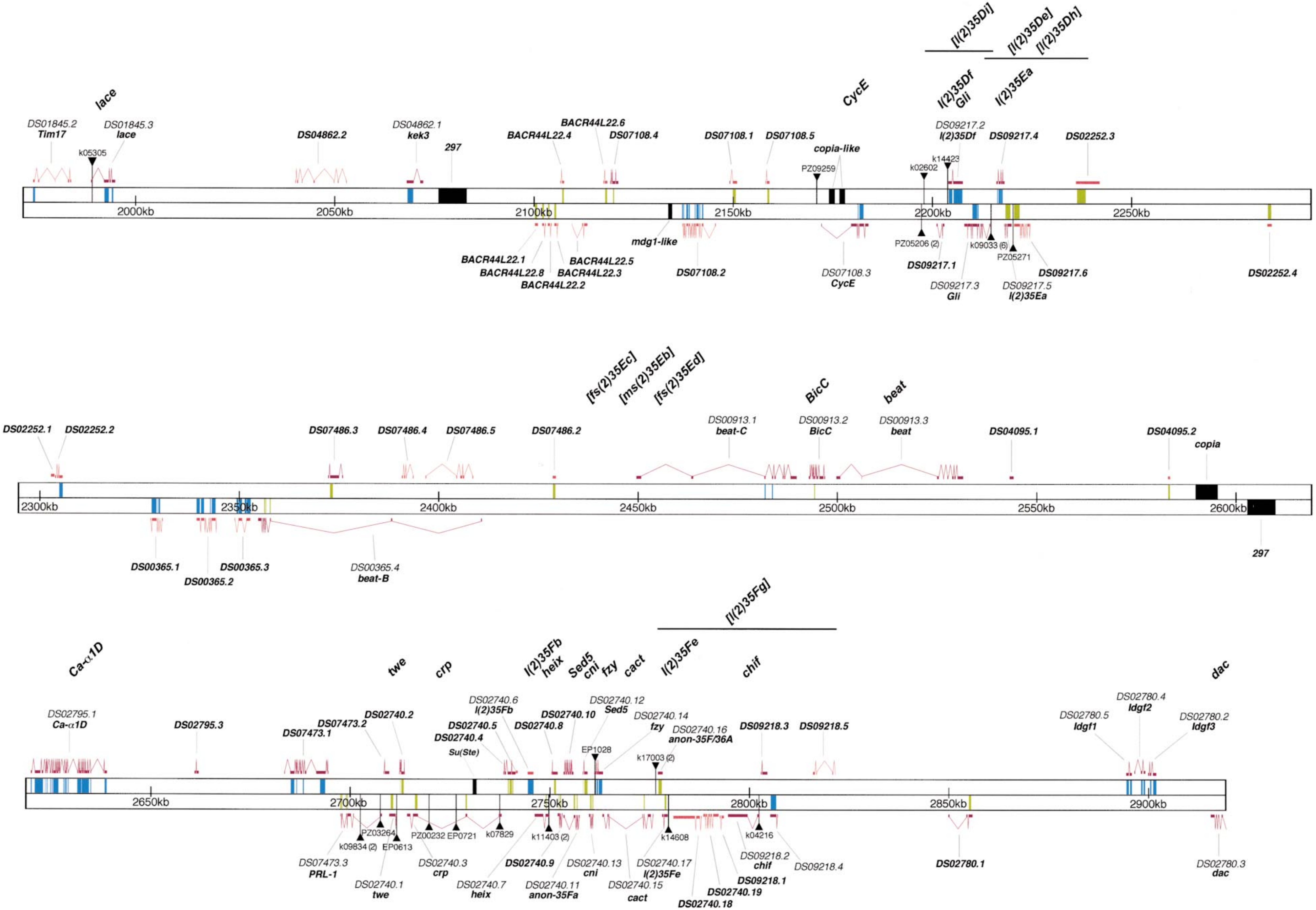
The data output from these various computational analyses is voluminous and requires intelligent filtering to remove redundant and irrelevant information before being passed to the human annotators. Moreover, the task of annotation is almost impossible without tools for the visualization of these data. An application, BLAST Output Parser (v. 01; BOP), was written (S. Lewis, unpublished results). BOP summarizes all automatically computed analysis data for an individual sequence into one file (*i.e.*, all output from the programs mentioned previously: BLAST, GENSCAN, etc.). This file is in

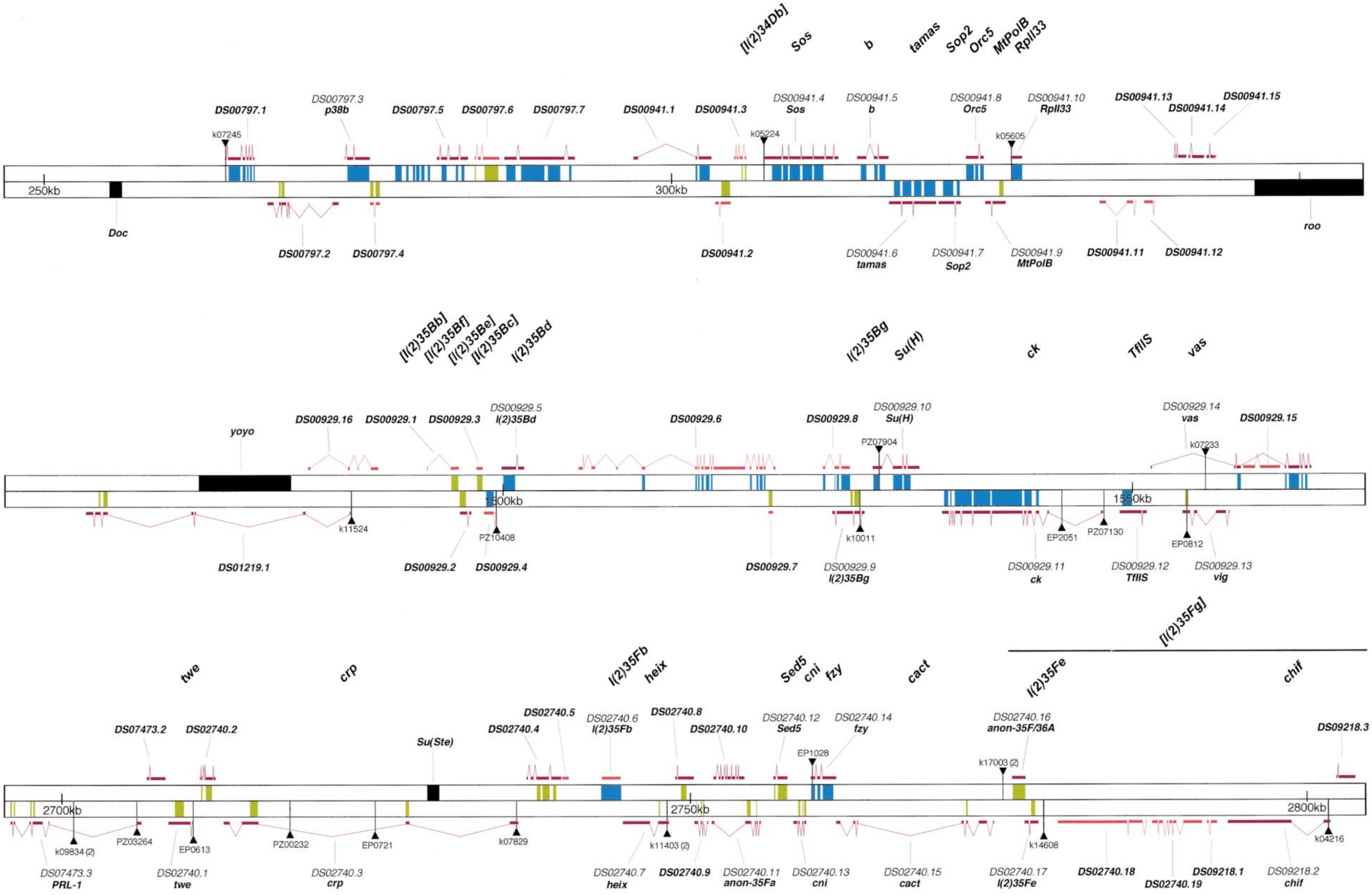
Figure 1.—A summary molecular map of the *Adh* region, covering 2.9 Mb of DNA. Genes located on the top of each map are transcribed from distal to proximal (with respect to the telomere of chromosome arm *2L*); those on the bottom are transcribed from proximal to distal. The gene symbols used in this figure are boldface type; if not the formal symbol then the latter is shown in a lighter font (formal symbols are abbreviated, their *BG:* prefix being omitted from Figures 1 and 2). *P*-element insertions are shown as triangles projecting to the molecular map. Red bars indicate transcribed regions, with intron-exon structures as predicted. Those in dark red are confirmed by a cDNA or were previously known; those in light red have only GENEFINDER or GENSCAN predictions (with cutoffs of 20 and 45, respectively). The blue and green boxes are BLASTX or TBLASTX matches detected using genomic DNA sequences from a GenBank submission (usually a single P1 or BAC clone) to search against sequences of other species in the databases. Similarities are shown in green for expectations between $P = 10^{-8}$ and $P = 10^{-50}$; blue for expectations of $P = 10^{-51}$ or lower. Once translations of predicted or known genes were used for BLASTP searches, some similarities that had not been detected using the nucleic acid sequence of the genomic clones were found. A summary of these BLASTP data is found in Table S2 (http://www.genetics.org/cgi/content/full/153/1/179/DC2). Transposable elements are indicated by black boxes and are named according to FlyBase. Genes defined genetically are shown above the map. Genes whose symbols are within square brackets are not tied to the map. These genes are indicated above a horizontal line when their order with respect to the genes below the line is not known. A scale in kilobases is shown; ∼1 cm = 10 kb.

Figure 2.—Enlarged views of the *Sos-RpII33*, *l(2)35Bb-vas*, and *twe-chif* regions. Symbols and conventions as in Figure 1. A scale in kilobases is shown; ∼3 cm = 10 kb.


kuz
[l(2)34Db]
Sos
b
tamas
Sop2
Orc5
MtPolB
RpII33
BACR48E02.2
BACR48E02.3
Fw
DS07660.3
kuz
DS01368.1
DS08249.2
DS08249.4
DS00797.3
p38b
DS00797.1
DS00797.5
DS00797.6
DS00797.7
DS00941.3
DS00941.1
DS00941.4
Sos
DS00941.5
b
DS00941.8
Orc5
DS00941.10
RpII33
PZ05337
k07245
k05224
k05605
50kb
100kb
150kb
200kb
250kb
300kb
EP2503
k01403 (6)
B4
DS07660.4
297
BACR48E02.1
DS07660.1
G
BACR48E02.4
DS08249.1
DS08249.3
DS08249.5
Doc
DS00797.2
DS00797.4
DS00941.2
DS00941.6
tamas
DS00941.7
Sop2
DS00941.9
MtPolB
Ance
[l(2)34Ec]
[j]
rk
l(2)34Fa
DS00941.13
DS00941.14
DS00941.15
DS08220.1
k10802
PZ06646 (2)
DS08220.3
Ance
DS00180.1
Acyp
DS00180.12
DS00180.5
DS00180.11
DS00180.9
DS00180.14
DS01514.3
k09909
DS01514.2
DS05899.1
DS05899.3
DS05899.5
DS05899.4
DS01759.1
DS01523.1
400kb
450kb
500kb
550kb
600kb
650kb
EP2171
rk[11p]
k00811
EP(2)2418
DS00941.11
DS00941.12
roo
DS08220.2
anon-34Ea
DS00180.2
DS00180.3
DS00180.7
DS00180.8
DS00180.10
DS00180.13
rk
DS01514.1
tRNA:gln
DS05899.2
l(2)34Fa
DS05899.7
DS05899.6
DS01759.2
smi35A
wb
[ms(2)34Fe]
[l(2)34Fc]
[l(2)34Fd]
l(2)35Aa
elB
[pu]
[elA]
DS03792.1
wb
PZ09437 (2)
DS03792.2
DS01068.1
DS01068.2
BG:DS01068.7
Rab14
k08712
DS01068.8
l(2)35Aa
DS01068.9
spel1
k07706 (3)
700kb
750kb
800kb
850kb
900kb
950kb
k11509
smi35A
k16716 (2)
DS01523.2
DS01523.3
smi35A
DS01068.10
DS01068.4
DS01068.5
DS01068.6
DS06238.1
ppk
DS06238.3
elB
DS06238.4
DS08340.1
Doc


noc
osp
Adh
DS04641.1
noc
DS04641.2
tRNA:pro
DS04641.3
tRNA:G3:35Ba
EP2000
DS01486.4
tRNA:pro
DS01486.2
tRNA:pro
DS01486.8
Adh
DS01486.9
Adhr
DS07721.1
DS07721.3
DS00810.1
DS00810.3
DS06874.1
DS06874.2
DS06874.3
DS06874.7
1000kb
1050kb
1100kb
1150kb
1200kb
1250kb
1300kb
k13218
rJ571
DS04641.4
tRNA:G3:35Bb
DS04641.5
tRNA:G3:35Bc
DS04641.6
tRNA:G3:35Bd
DS04641.7
tRNA:G3:35Be
DS04641.8
DS01486.1
DS01486.3
tRNA:pro
DS01486.7
osp
Jockey
roo
DS09219.1
DS07721.6
DS00810.2
DS06874.4
DS06874.6
copia
[ms(2)35Bi]
[l(2)35Bb]
[l(2)35Bf]
[l(2)35Be]
[l(2)35Bc]
l(2)35Bd
l(2)35Bg
Su(H)
ck
TfIIS
vas
stc
DS03323.1
DS00929.3
DS00929.1
DS00929.16
DS00929.5
l(2)35Bd
DS00929.6
DS00929.8
DS00929.10
Su(H)
PZ07904
DS00929.14
vas
k07233
DS00929.15
DS004929.1
DS04929.3
DS04929.4
stc
DS03192.3
yoyo
1350kb
1400kb
1450kb
1500kb
1550kb
1600kb
k11524
PZ10408
k10011
EP2051
PZ07130
EP0812
PZ05441
DS03431.1
DS03431.2
Mst 35Ba
DS03431.3
Mst 35Bb
Jockey
DS03144.1
DS01219.3
DS01219.1
DS00929.2
DS00929.4
DS00929.7
DS00929.9
l(2)35Bg
DS00929.11
ck
DS00929.12
TfIIS
DS00929.13
vig
DS03192.1
tRNA:leu
[rd]
[l(2)35Cc]
gft
ms(2)35Ci
esg
[l(2)35Cg]
wor
[l(2)35Ch]
sna
DS07295.4
DS07295.2
DS07295.5
DS07851.4
DS07851.5
DS07851.7
esg
DS03023.2
blood
1650kb
1700kb
1750kb
1800kb
1850kb
1900kb
1950kb
PZ06430
PZ02316
PZ07082 (21)
DS03192.2
DS03192.4
DS07295.1
DS07295.3
DS05639.1
DS007851.2
gft
DS07851.3
DS07851.8
DS07851.10
ms(2)35ci
DS07851.6
DS03023.4
DS03023.1
wor
DS01845.1
sna


[l(2)35Di]
[l(2)35De]
[l(2)35Dh]
lace
CycE
l(2)35Df
Gli
l(2)35Ea
DS01845.2
Tim17
DS01845.3
lace
k05305
DS04862.2
DS04862.1
kek3
297
BACR44L22.4
BACR44L22.6
DS07108.4
DS07108.1
DS07108.5
copia-like
PZ09259
k02602
k14423
DS09217.2
l(2)35Df
DS09217.4
DS02252.3
2000kb
2050kb
2100kb
2150kb
2200kb
2250kb
mdg1-like
BACR44L22.1
BACR44L22.8
BACR44L22.2
BACR44L22.3
BACR44L22.5
DS07108.2
DS07108.3
CycE
PZ05206 (2)
DS09217.1
k09033 (6)
PZ05271
DS09217.3
Gli
DS09217.5
l(2)35Ea
DS09217.6
DS02252.4
[fs(2)35Ec]
[ms(2)35Eb]
[fs(2)35Ed]
BicC
beat
DS02252.1
DS02252.2
DS07486.3
DS07486.4
DS07486.5
DS07486.2
DS00913.1
beat-C
DS00913.2
BicC
DS00913.3
beat
DS04095.1
DS04095.2
copia
2300kb
2350kb
2400kb
2450kb
2500kb
2550kb
2600kb
DS00365.1
DS00365.2
DS00365.3
DS00365.4
beat-B
297
[l(2)35Fg]
Ca-α1D
twe
crp
l(2)35Fb
heix
Sed5
cni
fzy
cact
l(2)35Fe
chif
dac
DS02795.1
Ca-α1D
DS02795.3
DS07473.1
DS07473.2
DS02740.2
DS02740.6
l(2)35Fb
DS02740.5
DS02740.4
Su(Ste)
DS02740.8
DS02740.10
DS02740.12
Sed5
EP1028
DS02740.14
fzy
DS02740.16
anon-35F/36A
k17003 (2)
DS09218.3
DS09218.5
DS02780.5
Idgf1
DS02780.4
Idgf2
DS02780.2
Idgf3
2650kb
2700kb
2750kb
2800kb
2850kb
2900kb
k09834 (2)
PZ03264
EP0613
PZ00232
EP0721
k07829
k11403 (2)
k14608
k04216
DS07473.3
PRL-1
DS02740.1
twe
DS02740.3
crp
DS02740.7
heix
DS02740.9
DS02740.11
anon-35Fa
DS02740.13
cni
DS02740.15
cact
DS02740.17
l(2)35Fe
DS02740.18
DS02740.19
DS09218.1
DS09218.2
chif
DS09218.4
DS02780.1
DS02780.3
dac



[l(2)34Db]
Sos
b
tamas
Sop2
Orc5
MtPolB
RpII33
DS00797.1
DS00797.3
p38b
DS00797.5
DS00797.6
DS00797.7
DS00941.1
DS00941.3
DS00941.4
Sos
DS00941.5
b
DS00941.8
Orc5
DS00941.10
RpII33
DS00941.13
DS00941.14
DS00941.15
k07245
k05224
k05605
250kb
300kb
Doc
DS00797.2
DS00797.4
DS00941.2
DS00941.6
tamas
DS00941.7
Sop2
DS00941.9
MtPolB
DS00941.11
DS00941.12
roo
[l(2)35Bb]
[l(2)35Bf]
[l(2)35Be]
[l(2)35Bc]
l(2)35Bd
l(2)35Bg
Su(H)
ck
TfIIS
vas
yoyo
DS00929.16
DS00929.1
DS00929.3
DS00929.5
l(2)35Bd
DS00929.6
DS00929.8
PZ07904
DS00929.10
Su(H)
DS00929.14
vas
k07233
DS00929.15
1500kb
1550kb
k11524
PZ10408
k10011
EP2051
PZ07130
EP0812
DS01219.1
DS00929.2
DS00929.4
DS00929.7
DS00929.9
l(2)35Bg
DS00929.11
ck
DS00929.12
TfIIS
DS00929.13
vig
[l(2)35Fg]
twe
crp
l(2)35Fb
heix
Sed5
cni
fzy
cact
l(2)35Fe
chif
DS07473.2
DS02740.2
Su(Ste)
DS02740.4
DS02740.5
DS02740.6
l(2)35Fb
DS02740.8
DS02740.10
DS02740.12
Sed5
EP1028
DS02740.14
fzy
k17003 (2)
DS02740.16
anon-35F/36A
DS09218.3
2700kb
2750kb
2800kb
k09834 (2)
PZ03264
EP0613
PZ00232
EP0721
k07829
k11403 (2)
k14608
k04216
DS07473.3
PRL-1
DS02740.1
twe
DS02740.3
crp
DS02740.7
heix
DS02740.9
DS02740.11
anon-35Fa
DS02740.13
cni
DS02740.15
cact
DS02740.17
l(2)35Fe
DS02740.18
DS02740.19
DS09218.1
DS09218.2
chif

XML syntax. BOP also removes as much of the "noise" as possible (*e.g.*, redundant matches, "shadow" matches on the noncoding strand, and matches to sequences of very biased base composition). These condensed data were then presented to the annotator in a graphical view (CloneCurator v. 0.1; S. Lewis, N. Harris, S. Misra and G. Helt, unpublished results).

CloneCurator was used to isolate individual genes from the clone sequences, based on expert evaluation of these analyses. CloneCurator allowed the annotator to compare results from different programs and to view the results using filters to determine a desired level of probability of prediction. The annotator used this visual summary to endorse a set of results as evidence, thereby generating a verified annotation. Annotations can be edited in CloneCurator and the annotators can add textual comments to any particular annotation, assign gene symbols, etc. This program was used to generate nucleic acid and amino acid FASTA files for each gene annotation. When a gene spanned more than one clone, manual intervention by an annotator was necessary to construct virtual mRNA sequences.

Open reading frames of predicted genes were validated using ORFfinder (v. 0.1; E. Frise, unpublished results) and all predicted proteins were then tested with BLASTP (v. 2.0a) with the options filter = SEG + XNU (unless the results are stated as being "unfiltered") against SWISS-PROT and SPTREMBL protein sets organized into nine taxonomic groups (Drosophila, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, other invertebrates, primates, rodents, other vertebrates, plants, and bacteria). Matches with an expectation below $P = 10^{-7}$ were ignored.

Protein domains and motifs were analyzed against the PROSITE (release 15.0; Hofmann *et al.* 1999) and PFAM (v. 2.1.1; Sonnhamer *et al.* 1997; Bateman *et al.* 1999) databases using the programs PPSEARCH [a Unix implementation of MacPattern at http://www2.ebi.ac.uk/services.html/ (Fuchs 1994)] and HAMMER2.1 (Eddy 1998). PROSITE output was filtered using EMOTIF (Nevill-Manning *et al.* 1998) at the European Bioinformatics Institute (EBI). The SAPS program (version of July 23, 1993; Brendel *et al.* 1992) was run from the EBI server (http://www2.ebi.ac.uk/SAPS/) to analyze various compositional features of predicted protein sequences. The PSORTII suite of programs (Horton and Nakai 1997), trained on the proteins of *S. cerevisiae*, was used to predict the subcellular localization of proteins. Sequence alignments were generated using CLUSTALW (Higgins *et al.* 1996) from the European Bioinformatics Institute server (http://www2.ebi.ac.uk/services.html/).

The output from the various sequence analysis programs is archived on FlyBase as FlyBase-Annotation files linked to the sequenced clones. Version 1 of these files includes the analyses used for this article. Subsequent versions will result from reanalysis of the sequence data.

Nomenclature: All genes are named according to the conventions agreed between the Berkeley and European *Drosophila* Genome Projects and FlyBase (http://flybase.bio.indiana.edu/docs/nomenclature). Each gene is given a unique name composed of three parts: a prefix (*BG* for genes defined by the Berkeley Project, *EG* for those defined by the European Project), followed by a clone name and an integer. The clone name is that of the clone on which the gene was first defined (regardless of whether or not the gene overlaps more than one clone). The final integer is simply a serial number, and does not imply the order of a gene within a clone. An example is *BG:DS09218.6*, the sixth gene annotated on P1 clone DS09218. If a gene was already known to FlyBase, then a formal name is still assigned but will be treated by FlyBase as a synonym of the established name.

All genes known to FlyBase are named by those names and symbols declared by FlyBase as valid. In addition, the historical names of the lethals identified by the genetic analysis of the *Adh* region are given.

Availability of data and materials: The DNA sequence of the *Adh* region is made available for file transfer protocol (ftp) and searching (using BLAST) at http://www.fruitfly.org/data/genomic_fasta/Adh_and_cactus. All sequence data from genomic clones, ESTs, cDNAs, and *P*-element flanking regions are deposited in GenBank. Supplementary tables of data, cited in this article as Tables S1, S2, and S3, are available from http://www.genetics.org/supplemental/. Accession numbers for the genomic sequences are given in Table 2, for *P*-element flanking regions in Table S1 (http://www.genetics.org/cgi/content/full/153/1/179/DC1), and for cDNAs and ESTs in Table S2 (http://www.genetics.org/cgi/content/full/153/1/179/DC2). P1 clones are available from laboratories listed on FlyBase. cDNA clones are available from Research Genetics (Huntsville, AL) or from Genome Systems (St. Louis, MO). BAC clones (library RPCI-98) are available from Dr. P. de Jong (Roswell Park Cancer Institute, Buffalo, NY). *P*-element alleles are available from the Bloomington and Szeged *Drosophila* Stock Centers or from the Berkeley *Drosophila* Genome Project (BDGP). The annotated sequences can be viewed through FlyBase as CloneCurator reports.

RESULTS AND DISCUSSION

The physical map and sequence of the *Adh* region: The physical map of the *Adh* region was assembled and sequenced from P1 and BAC as described in materials and methods. The P1 clones formed three contigs, one of 1,940,896 bp, one of 798,089 bp, and the third, a single P1 clone. The gap between the 1.9-Mb and 0.79-Mb contigs could not be closed in P1 clones, but was, however, readily closed by screening the BAC library; it was found to be 43,803 bp in length. A BAC clone also linked the isolated P1 clone (DS07660) to the distal end of the 1.9-Mb contig. This gap was 35,162 bp in length. The total length of sequence studied is 2,919,020 bp. A summary of the interpretation of this sequence is given in Figure 1, with an expanded view of three selected regions in Figure 2.

General features of the sequence: The overall base composition of the sequence is 40.82% G + C, to be compared to the figure of 43% for the genome as a whole (Laird and McCarthy 1969). The G + C contents of functionally different regions of the sequence, protein-coding regions, introns, and intergenic spacer are 49.7, 38.7, and 39.6%, respectively (intergenic regions may well be overestimated in size, because the gene prediction programs will have missed 5′ exons distant from the body of a gene unless full-length cDNAs were available). The average number of exons per gene is 4.4, but this figure must be treated with caution for the reasons just mentioned.

Gene prediction in the *Adh* region: A primary objective of the sequence analysis was to identify genes, both protein coding and others (*e.g.*, tRNA), in the 2.9 Mb of sequenced DNA. We predict the existence of 229, of which 218 are predicted to be protein coding and 11 tRNA coding (Figure 1). The bases for the predictions are summarized in Table S2 (http://www.genetics.org/cgi/content/full/153/1/179/DC2). Forty-one of the protein-coding genes are predicted only on the basis of a high score with a gene-finding program; of these, 16 have both GENSCAN and GENEFINDER predictions (above the thresholds we used), 2 have only GENEFINDER predictions, and 23 only GENSCAN predictions. All of the other protein-coding genes are predicted by either (or both) sequence similarities (a BLAST score of $P = <10^{-7}$; 156, 71%) or a match with a Drosophila EST, cDNA, or genomic sequence (110, 52% of protein-coding genes). (Seventeen more genes had matches to Drosophila ESTs, but these matches were clearly due to the ESTs being derived from genes encoding similar sequences, *i.e.*, from paralogous genes.)

It is important to get an estimate of the false-negative and false-positive frequencies of prediction. A GENSCAN threshold of 45 fails to predict 22 protein-coding genes predicted by other means (or known prior to this work). Of these 22, 10 have EST matches and 3 were known prior to this analysis (*Mst35Ba*, *Mst35Bb*, and *cni*). Lowering the threshold for GENSCAN to 30 would include 8 of these 22 false negatives, but this would also predict a further 25 protein-coding genes in this region, none of which would have any other support. The GENEFINDER program, at a threshold of 20, fails to predict 56 of the protein-coding genes. Of these false negatives, 35 have support from experimental data and 21 have support from GENSCAN predictions [Table S2 (http://www.genetics.org/cgi/content/full/153/1/179/DC2)]. One feature of GENSCAN that we have noticed is that its scores tend to be low in regions of very high gene density.

ESTs and cDNA sequences of genes in the *Adh* region: Even the best computational methods are imperfect in their ability to determine the intron-exon structures of genes from genomic sequence alone. Moreover, because such methods rely on information from codon usage and the maintenance of open reading frames, they are inherently unable to predict the presence of introns in 5′ or 3′ untranslated regions or to predict the transcriptional start sites. For these reasons it is necessary to isolate and sequence cDNAs (or RT-PCR products). We have used sequence matches between the genomic sequence and 5′ ESTs as a rapid way of identifying cDNAs for sequencing [see materials and methods; Table S2 (http://www.genetics.org/cgi/content/full/153/1/179/DC2)]. cDNAs corresponding to 95 genes were identified by matches to ESTs (44% of known or predicted protein-coding genes) at a time when the total number of Drosophila ESTs available was 53,000.

Of the 68 protein-coding genes for which there was some prior knowledge (*i.e.*, both genetic and molecular data or molecular data alone), 50 (74%) have ESTs; of the 150 genes that are newly discovered, only 44 (29%) have ESTs. This is a rather surprising result. It may indicate either a bias in the sample of genes that had already been studied or an overprediction of new genes, or it may be a biologically interesting result (see below).

***P*-element hits:** Several collections of lethal *P* elements were screened against deletions that, in sum, covered the entire *Adh* region (see Spradling *et al.* 1995, 1999). We have also analyzed genetically *P* elements from these collections that had not been recovered in the screens for lethals or semilethals, but which were found to map to the region by *in situ* hybridization to polytene chromosomes or by a sequence match of the sequences flanking the *P*-element insertion (Spradling *et al.* 1999). Similarly, sequences flanking 2300 insertions of the *EP* element (Rørth *et al.* 1998) were determined (J. Rehm and G. Rubin, unpublished data) and used to identify 24 *EP* insertions in this region. From these screens, and from those identified by others, 181 independent *P*-element insertions in 43 genes have been identified [Tables 1 and S1(http://www.genetics.org/cgi/content/full/153/1/179/DC1)]. *P*-element insertions in 35 genes give a lethal, or semilethal, sterile, or visible phenotype. In the remaining eight genes all known insertions are without obvious phenotypic effect.

Gene density in the *Adh* region: Of the 229 genes, 218 are protein coding and 11 are tRNAs. The average gene density for protein-coding genes is one per 13.4 kb. The average size of the genes, as estimated both from computational analysis and the "full"-length cDNAs, is 5.5 kb (from ATG to terminator, including introns). The average gene density of one gene per 13.4 kb hides enormous variation in density. Some regions are very dense, with genes being separated by only a few hundreds of base pairs; others are, by comparison, very gene poor (see Figures 1 and 2).

There are few studies of long genomic sequences of Drosophila that we can use for comparison with the *Adh* region. Preliminary analyses of 2 Mb of genomic sequence from region 1–3 of the *X* chromosome give a gene density of one gene per 8 kb (T. Benos and M. Ashburner, unpublished analyses of European *Drosophila* Genome Project data). In the 338-kb *bithorax* region there are 13 known or predicted genes (1 per 24 kb), but 3 of these (*Ubx*, *abd-A*, and *Abd-B*) are exceptionally large (22 to 78 kb for their coding regions alone). In the *Antp* region Celniker *et al.* (S. Celniker, B. Pfeiffer, J. Knafels, C. Mayeda, C. Martin and M. Palazzolo, unpublished data) have identified 26 protein-coding genes in 430 kb, a density of 1 gene per 16.5 kb. Maleszka *et al.* (1998) predicted 12 genes within one 67-kb P1 clone from the base of the *X* chromosome (1 gene per 5.6 kb).

Transcriptional bias: The number of genes transcribed from each DNA strand is approximately equal (121 *vs.* 108). In very gene-dense regions there is a strong tendency for the direction of transcription to

TABLE 3

Selected regions of the gemone of *D. melanogaster* subjected to "saturation" genetic analysis for lethal complementation groups, showing the average ratio of lethal loci to polytene chromosome bands

Region	Band no.	Lethal groups	Average no. mutations/ lethal group	Lethals/band	Reference[a]
Adh	69	55	10.8	0.81	1
z-w	16	12	9.7	0.75	2
ry	24	20	7.6	0.83	3
Ddc	10	15	16.9	1.80	4
dpp	13	13	6.4	1.0	5
pk	28	20	3.9	0.71	6
kar	9	4	10.5	0.44	7
v	12	7	10.5	0.58	8
tra	17	15	4.0	0.88	9
Total	198	161	—	0.81	—

[a] 1, this study (includes predicted lethal loci; see Table 1); 2, Judd *et al.* (1972); 3, Hilliker *et al.* (1981); 4, Stathakis *et al.* (1995); 5, Littleton and Bellen (1994); 6, Heitzler *et al.* (1993); 7, Gausz *et al.* (1979); 8, Kozlova *et al.* (1994); 9, Belote *et al.* (1990).

alternate (see Figure 1); overall, however, the pattern of transcriptional direction appears to be random. This was tested by expressing the pattern as a binary string and attempting to compress it using the Lempel-Ziv compression algorithm (Ziv and Lempel 1977). The string did not compress any better than did 1000 randomly generated strings of the same length.

Estimates of total gene number in Drosophila: Any estimate of total gene number, based on the analysis of the *Adh* region, depends on this region being "typical" of the genome as a whole, with respect to the number of genes. This is a difficult question to answer with any rigor. Genetically, there are no indications that the *Adh* region is atypical. The number of genes discovered by genetic analysis is, given the number of polytene chromosome bands included, very similar to that in other well-studied regions. Classical "saturation" studies give a ratio of lethal complementation groups to polytene chromosomes bands of ∼0.84 (Table 3); for the *Adh* region this ratio is 0.81.

Our estimates of the total gene number rely on estimates of the total DNA content of *D. melanogaster*. This has been independently estimated to be 170 Mb by Rudkin (1972; and cited in Kavenoff and Zimm 1973), using UV microspectrophotometry of diploid ganglion cells by Rasch *et al.* (1971), by Feulgen microspectrophotometry of sperm and haemocyte cells, and by Kavenoff and Zimm (1973) from the kinetics of relaxation of whole chromosome-length DNA molecules. The kinetics of reassociation of denatured DNA gave a slightly lower estimate (Laird 1971). Of this 170 Mb of DNA, some 21% is estimated to be low-complexity satellite sequence (Lohe and Brutlag 1987) and 12% transposable elements and other repeated sequences, such as the histone and rRNA genes (Laird and McCarthy 1968). This gives an estimate of ∼115 Mb of "unique" DNA sequence.

Simple arithmetic, 115 Mb/13.4 kb, gives an estimate of 8600 protein-coding genes for the Drosophila genome as a whole. This is a remarkably low number, being less than half as much again as the yeast *S. cerevisiae* (6000; Mewes *et al.* 1997) and less than half the number now estimated for *Caenorhabditis elegans* (19,090; The *C. elegans* Sequencing Consortium 1998). An independent estimate can be made, knowing that the sequenced region covers 69 polytene chromosome bands, an average of 42 kb/band plus its adjacent interband [rather higher than Sorsa's estimate of 21.6 kb/band (Sorsa 1988)]. The total band number is estimated to be 5160 (V. Sorsa, quoted in Ashburner 1989). In terms of band number, therefore, the *Adh* region is 1.34% of the total. If the density of genes per band in this region is typical of the genome as a whole, then this leads to an estimate of 16,975 genes. Our two estimates of the total gene number in *D. melanogaster*, 8600 and 16,975, bracket the estimate of 12,000 by Miklos and Rubin (1996), based on the sizes of 276 individual genes.

Local duplications of genes: A number of genes in Drosophila have been found to exist as locally duplicated gene pairs. Members of a pair may be functionally distinct (*e.g.*, *en*, *inv*) or functionally redundant (*e.g.*, *gsb-d*, *gsb-p*; *ph-d*, *ph-p*). The most obvious model for the origin of gene pairs is unequal recombination (Sturtevant 1925; Ingram 1961; Baglioni 1963; Smithies *et al.* 1962) followed by sequence divergence.

In this chromosome region we have identified at least 12 (protein-coding) gene repeats. One had already been identified, first in *Drosophila pseudoobscura* (Schaeffer and Aquadro 1987), *i.e.*, *Adh* and *Adhr*, genes just 300 bp apart that have protein products only 33% identical in sequence, yet with a conserved position of introns. Remarkably, *Adhr* is only transcribed as a dicistronic transcript with *Adh* (Brogna and Ashburner 1997). The second gene repeat is a triplication of three zinc finger domain transcription factors, *escargot*, *worniu*, and *snail*, within 150 kb. The proteins encoded by these genes show 31–37% pairwise identity. Interestingly, although each of these is required for viability, there is some residual functional redundancy between at least *esg* and *sna* (see appendix). The third example is *BG:DS01514.2* and *BG:DS05899.1*, two genes 7.5 kb apart that encode protein products 43% identical in sequence; these proteins show similarity to mouse long-chain fatty acid coenzyme-A ligase. *Mst35Ba* and *Mst35Bb* are a tandem pair of genes encoding protamine-like proteins characterized by Russell and Kaiser (1993). These proteins are 91% identical over their common region (that of *Mst35Bb* is longer by 25 amino acids than that of *Mst35Ba*). At the nucleic acid level the duplication extends over ~1 kb.

Five genes, closely clustered in the region between *RpII33* and *Ance*, show between 30 and 37% amino acid sequence similarities. These are *BG:DS00941.11–BG:DS00941.15*, genes whose proteins are about the same size but all lack any sequence matches. *BG:DS00180.7–BG:DS00180.10*, *BG:DS00180.12*, and *BG:DS00180.14* are six genes all with epidermal growth factor (EGF) domains clustered within a few tens of kilobases just distal to *rk*. Their sequence similarities are not high, but are evidence of ancient duplications.

In the region between the *lace* and *CycE* genes there are six predicted genes within 21 kb, each encoding a protein of the astacin subfamily of Zn-metalloproteases (Barrett *et al.* 1998; *BG:BACR44L22.1–BG:BACR44L22.4*, *BG:BACR44L22.6*, and *BG:BACR44L22.8*). The predicted protein sequences of these genes are between 29 and 64% identical. There are two clusters of genes encoding proteins predicted to be serine proteases. One is of two genes within 14.8 kb and showing 45% pairwise similarity (*BG:DS06874.4* and *BG:DS06874.6*); the other is a pair of genes within 10.2 kb showing 35% sequence similarity (*BG:DS07108.1* and *BG:DS07108.5*). Right at the proximal margin of the region sequenced are three genes encoding proteins identified by Kawamura *et al.* (1999) as imaginal disc growth factors (see below). These genes show 51–55% pairwise similarity in sequence and are within 7.7 kb (*Idgf1*, *Idgf2*, and *Idgf3*). Interestingly, there is evidence for a tandem triplication of chitinase genes, which these resemble, in mosquitoes (De La Vega *et al.* 1998). A further triplication is exemplified by *beat* and two similar genes, *beat-B* and *beat-C*, first discovered in this sequence by T. Pipes (personal communication). These three genes are not contiguous, but are clustered within 200 kb. The proteins predicted for *beat-B* and *beat-C* are 51 and 46% identical, respectively, to that of *beat*. The three genes have a similar structure. The final example of duplicate genes is that of *noc* and *BG:DS06238.3*, a gene some 100 kb distal, which we suggest is *elB* (see below). These two genes encode Zn-finger proteins with 27% amino acid identity.

The 38 genes in the 34C-36A region that appear to be members of tandem series represent 17% of the total number of protein-coding genes. This is a minimum estimate, because a BLASTP search of all 218 known and predicted protein sequences against themselves identifies other potential duplications, which require further study. Many of these duplications are very old, as judged by the sequence similarities between members of a set. Tandem series of genes are also a feature of *C. elegans* (The *C. elegans* Sequencing Consortium 1998; The *C. elegans* Genome Sequencing Project 1999) and *Arabidopsis thaliana* (Bevan *et al.* 1998). The fraction of genes included in tandem sets of two or more (18%) is about the same as that found in the *Adh* region (Jones 1999). One possible reason why *C. elegans* appears to have more genes than *D. melanogaster* would be that these local tandem arrays are, on average, larger in *C. elegans*. The data available so far do not support this suggestion.

Genes within genes: The first example of a gene known to be entirely included within another gene was that of a pupal cuticle protein gene (*Pcp*) fully encoded within an intron of *ade3* (Henikoff *et al.* 1986). Since then, >30 examples have been discovered (data from FlyBase) and in the majority of cases (25/32) the included gene is transcribed from the opposite strand of the including gene. In the *Adh* region we have identified 17 examples of nested genes, 12/17 following the majority rule of antiparallel transcription.

The inclusion of *Adh* within *osp* was first suggested by genetic data, because *osp* aberrations mapped to either side of *Adh* (Chia *et al.* 1985; see below). This suggestion, and the inclusion of *Adhr* in the same intron, was confirmed by molecular analysis (McNabb *et al.* 1996) and is proven here by the comparison of the sequence of a full-length *osp* cDNA with the genomic sequence (see below). Two other predicted genes are within *osp*: *BG:DS07721.1* and *BG:DS09219.1*.

An open reading frame in the 5′ intron of *vasa* (*vig*, for *v*asa *i*ntronic *g*ene) was first identified by K. Edwards (personal communication) by a comparison of sequences from *D. grimshawi* with those from this project. There is another CDS within *vasa*: *BG:DS00929.15* in the long third intron, first identified as a ubiquitous transcript from RNA blots with genomic DNA by P. Lasko (personal communication; see Styhler *et al.* 1998). The other examples of putative included genes are *BG:BACR48E02.1*, *BG:BACR48E02.2*, and *BG:BACR48E02.3*, all included within the second intron of *B4*; *BG:DS07486.3*, *BG:DS07486.4*, and *BG:DS07486.5* included within introns of *beat-B*, the former in intron 1 and the latter two in intron 2; *BG:DS03792.2* is within *wb*; *BG:DS03192.4* is within *BG:DS03192.2*;

BG:DS07295.4 is within *BG:DS07295.1*; *BG:DS07660.1* is within *kuz*; and *BG:DS01514.1* is within *BG:DS01514.3*.

The phenotypes of overlapping and contiguous deletions—the search for more genes: We have evidence that the genetic screens failed to recover mutations at loci expected to have scorable phenotypes—the failure to recover any alleles of *beat* is an example (see appendix). One new lethal locus (*l(2)35Fg*) was discovered when the chromosome *2 P* elements were systematically screened. One further genetic technique to discover genes is to systematically screen hetetozygotes between two overlapping deletions. We have made transheterozygotes between all possible pairs of deletions, which, by genetic criteria, abut, *i.e.*, the distal end of one and the proximal end of another are located between the same pair of genes identified by mutant alleles. These pairs of deletions may or may not physically overlap.

Pairwise combinations (836) have been made and the genotypes scored for viability, male and female fertility, and obvious visible phenotypes. Although these phenotypes could be the result of the additive effects of haplo-insufficiency, we have predicted the existence of four lethal loci from these data, two loci required for male fertility and two loci required for female fertility (each "locus" could include more than one gene, of course). A variation on this protocol for the discovery of mutant phenotypes is to test combinations of deletions that are known to overlap by only one gene with a mutant phenotype in the presence of a transgene that is known independently to rescue the mutant phenotype. If the transgene rescues the deficiency heterozygote to phenotypic normality, then we can conclude that no other genes capable of giving a mutant phenotype are located in the deleted interval; and if not, then we can conclude the existence of a previously unsuspected locus.

Overlapping *Ance*$^{-}$ deletions are lethal, which is expected, since *Ance* itself is a vital gene. There is, however, evidence for another lethal near *Ance*, because the lethality of some, but not all, overlapping deletion pairs can be rescued by a 16.5-kb transformant that includes both *Ance* and *anon-34Ea* (carried on *P{RACE}*). *l(2)34Ec* is predicted on the basis of the failure of this transformant to rescue the lethality of, *e.g.*, *Df(2L)SR407*/*Df(2L)b82a1*. This predicted gene is not in the overlap of, *e.g.*, *Df(2L)SR407*/*Df(2L)b74c6*.

The existence of *ms(2)35Bi*, between the 5′ exons of *osp* and *l(2)35Bb*, is predicted on the basis of viable, but male-sterile, overlapping deletion heterozygotes (see appendix). *l(2)35Cc* is predicted on the basis of the recessive lethality of *Df(2L)rd9* (Ashburner *et al.* 1990). *rd9* is lethal with deletions of *rd*; all five other known alleles of *rd* are hemizygous viable. The existence of *l(2)35Cc* is confirmed by the complementation behavior of deletions generated from *gft*^{PZ06430} by male recombination. Of nine deletions, one extended distally and was *rd*$^{+}$ but lethal with *Df(2L)rd9* and *gft*; the other eight extended proximally from *gft* to include *ms(2)35Ci*.

The region between *esg* and *sna* is, genetically, rather complex. From the phenotypes of overlapping deletions Ashburner *et al.* (1990) identified a region that, when homozygously deleted, can result in either lethality or an absence of the halteres. These phenotypes are separable; *e.g.*, the *Df(2L)osp38*/*Df(2L)TE35D-22* heterozygote is viable and lacks halteres, but *Df(2L)osp18*/*Df(2L)TE35D-22* is lethal. Both map between *esg* and *worniu*. The lethal is here named *l(2)35Cg*. There is another predicted lethal in this region, simply called *l* by Ashburner *et al.* (1990; Figure 2). It (*l(2)35Ch*) is predicted from the lethality of, *e.g.*, *Df(2L)el20* when heterozygous with *Df(2L)Scorv25*. There is only one gene prediction in the *esg*-*worniu* interval; this is *BG:DS03023.4*.

fs(2)35Ec is inferred from the sterility of *Df(2L)RA5* females heterozygous with 18 different deletions, *e.g.*, *Df(2L)TE35D-3*. The existence of *fs(2)35Ed* is suggested by the sterility of *Df(2L)RM5*/*Df(2L)TE35D-2* females and of four similar genotypes; this gene may correspond to *beat-C*. *ms(2)35Eb* is inferred from the male sterility of the heterozygote *Df(2L)RA5*/*Df(2L)TE35D-14*. The predicted female steriles, *fs(2)35Ec* and *fs(2)35Ed*, are tentative; we are concerned that these phenotypes may simply result from haplo-insufficiency, particularly for *BicC*.

There are several regions that are homozygous viable when deleted. We estimate that the longest of these, the overlap of *Df(2L)A178* and *Df(2L)A446*, is 190 kb. This overlap deletes or disrupts four known genes (*noc*, *Adh*, *Adhr*, and *osp*), eight tRNA genes, and five predicted protein-encoding genes in the *noc*-*BG:DS07721.3* interval.

The structure and function of gene products: We have used three computational techniques to infer structural and functional attributes of the products of the genes predicted for this chromosome region. These are searches for protein motifs or domains using the PFAM and PROSITE databases, BLASTP similarities of the predicted open reading frames with proteins in the SWISSPROT and SPTREMBL databases, and some analysis of protein features using the PSORT and SAPS programs (see materials and methods). In general, we have been rather conservative in making these inferences, as we have for gene prediction in general. These functional inferences are summarized in Table S3 (http://www.genetics.org/cgi/content/full/153/1/179/DC3), using a classification now being developed by the Gene Ontology Consortium (FlyBase, Mouse Genome Informatics and the Saccharomyces Genome Database; Go 1999). Of the 218 known or predicted protein-coding genes, we know, from previous work by others, or have inferred, the function of less than half (91, 42%). Of these, 41 are obviously enzymes and 18 are predicted to be proteases; the rest cover the functional spectrum from structural proteins (*e.g.*, cuticle protein) to growth factors and transporters. From our analysis of protein motifs we predict that 16 of the proteins are

DNA or RNA binding; the PSORT analysis predicts that 82 are nuclear localized, but this may well be an overestimate. There are some features of the domain analysis that deserve further study: the cluster of six genes (*BG:DS00180.10* and neighbors) whose products are predicted to have EGF domains in particular.

Evolutionary conservation: Of the 156 known or predicted protein-coding genes, 72% have clear matches with those in other organisms [summarized in Table S2 (http://www.genetics.org/cgi/content/full/153/1/179/DC2)]. Of these, 120 have matches to the sequences of *C. elegans*, 69 to the sequence of *S. cerevisiae*, 35 to sequences of *A. thaliana*, 114 to sequences from rodents (nearly all mouse, with a few rat), 125 to human sequences, and 128 to rodent + human sequences. Thirty proteins have matches in yeast, *C. elegans*, Arabidopsis, and rodents + human, and 55 in yeast, *C. elegans*, and rodents + human. With the exception of *S. cerevisiae* and *C. elegans* (whose genomes are entirely sequenced, or almost so) these numbers reflect the available sequence data, although, overall, they are an impressive witness to the conservation of protein sequence across very different taxa. These sequence similarities are, of course, very useful for making functional inferences about new Drosophila genes; they must, however, be treated with some caution as the evolution of function and sequence may not be as tightly linked as is sometimes believed. We see evidence for this in the genes of this region; *e.g.*, the fact that the three genes we first identified by their sequence characteristics as chitinases are in fact secreted imaginal disc growth factors, as has been shown experimentally (Kawamura *et al.* 1999). The inferences we have made are only hypotheses that demand experimental verification or falsification.

In addition to sequence similarities between genes in this chromosome region and sequences from other taxa, 49 of the predicted or known protein-coding genes have significant database matches outside the *Adh* region to the known protein universe of Drosophila. This is from a sample of only 2000 or so proteins, <15% of the expected total. The conclusion, which is no great surprise, is that nearly all proteins of Drosophila will be members of protein sequence families. In some cases the similarities in sequence between different proteins are very striking, *e.g.*, the two "stress-activated" mitogen activated protein (MAP) kinases *p38b* and *Mpk2* are 77% identical in sequence (see appendix). There is no obvious clustering of the genes that are paralogs of genes in the *Adh* region; this would have been evidence of large-scale genomic duplications, such as are found in *S. cerevisiae* (Wolfe and Shields 1997).

Correspondence between known genes and the sequence: One of the major objectives of this study was to identify the 73 genes known or predicted from the genetic analyses on the sequence and, if possible, to infer their function. For those that had been sequenced previously their identification was straightforward. Others have been identified by mapping to the sequence the sites of insertion of *P*-element alleles and by correlating the genetic and sequence maps. Forty-nine of these 73 genes have been identified on the sequence [see Figure 1 and Table S2 (http://www.genetics.org/cgi/content/full/153/1/179/DC2)]. For the remaining 24, candidate sequences can be identified, but no firm correlation can be made on the available data. Detailed consideration of these 49 genes and others of interest identified on the sequence is given in the appendix.

Genes with phenotypes are more likely to be conserved: Genes that can mutate to an observable phenotype are far more conserved than those that cannot. The data are shown in Table 4. We compare the sequence similarities between known and predicted proteins in two groups: the first is of all 218 proteins, the second just that subset of 49 encoded by genes for which we have phenotypically detectable mutant alleles. Even at a BLASTP threshold of $P = 10^{-50}$, 63% of the 49 genes with phenotypes (and known sequences) have sequence similarities in other taxa, compared to only 31% for the total sample of 218 genes. This difference is also observed if one only considers the comparisons to individual species, such as *C. elegans* and *S. cerevisiae*, whose genomes are completely sequenced; this argues that the observation cannot be due to an ascertainment bias.

We know, or predict from genetic data, that 73 out of 218 genes have mutant phenotypes. If we assume that the 24 genes that we have not yet managed to tie to the sequence are as conserved as the 49 that we have, then we can calculate the expected properties of the total sets of genes with and without mutant phenotypes. For example, we can predict 46/73 will have BLASTP hits to other species at an expectation of $P = 10^{-50}$. Because there are only 67 hits to other species from the total of 218 genes (at this cutoff) we can conclude that 63% of the genes with mutant phenotypes are conserved, but only 14% (21/(218-73)) of the genes without detectable mutant phenotypes. If we raise the BLASTP cutoff to $P = 10^{-100}$, then the numbers are even more striking: 37 and 2%, respectively, for genes of the two classes.

We realize that this analysis has its limitations. The distinction between genes with and without discernible mutant phenotypes is not hard and fast, but we point out that the great majority of mutant phenotypes known in this chromosome region are very obvious, *i.e.*, lethality, sterility, or marked changes to adult morphology. We can, in addition, have reasonable confidence that mutations have been detected in nearly all of the genes in this region that can mutate to these phenotypes.

Conserved genes are more highly expressed: Genes known previous to this analysis are far more likely to have ESTs than those newly discovered (see above). We were concerned that this could indicate an overoptimism in predicting new genes. Yet the analysis of Table 4 shows that this cannot be so, or at least it cannot be the entire reason. Genes with BLAST similarities with

TABLE 4

A comparison of the sequence similarities between genes with known mutant phenotypes and those without

P value	Other species	Vertebrates	*C. elegans*	*S. cerevisiae*	Plants	Bacteria	Drosophila	All
% of the 218 predicted genes in the *Adh* region with BLAST scores better than the indicated *P* value when compared to the indicated subsets of GenBank[a]								
<e-7	66 (51)	57 (51)	55 (53)	31 (66)	24 (68)	30 (51)	47 (48)	71 (48)
<e-20	51 (55)	45 (55)	37 (55)	19 (78)	17 (64)	12 (60)	36 (51)	58 (53)
<e-50	31 (60)	27 (64)	18 (67)	9 (80)	8 (82)	3 (50)	25 (63)	41 (60)
<e-100	14 (87)	13 (93)	8 (83)	3 (100)	3 (83)	0	17 (81)	23 (80)
% of 49 genes known to display loss-of-function phenotypes, with BLAST scores better than the indicated *P* value when compared to the indicated subsets of GenBank[a]								
<e-7	90 (80)	84 (78)	76 (78)	55 (81)	39 (84)	43 (71)	78 (71)	94 (80)
<e-20	82 (80)	76 (78)	64 (77)	37 (100)	27 (85)	18 (89)	76 (70)	94 (80)
<e-50	63 (77)	61 (77)	37 (94)	22 (100)	14 (100)	2 (100)	67 (73)	84 (77)
<e-100	37 (100)	37 (100)	20 (100)	10 (100)	6 (100)	6	53 (81)	65 (84)
% of 145 genes predicted to lack loss-of-function phenotypes, with BLAST scores better than the indicated *P* value when compared to the indicated subsets of GenBank[a]								
<e-7	54 (27)	44 (25)	45 (32)	19 (41)	17 (48)	23 (32)	32 (20)	59 (26)
<e-20	35 (25)	29 (24)	23 (24)	10 (36)	12 (41)	10 (29)	17 (4)	40 (26)
<e-50	14 (19)	9 (23)	8 (8)	3 (0)	5 (57)	3 (40)	3 (0)	19 (25)
<e-100	2 (0)	0	2 (0)	0	1 (50)	0	0	2 (33)

To calculate the expected percentage of the 145 genes that did not have loss-of-function phenotypes (218 total genes—73 with such phenotypes) we made the assumption that the 24 genes with phenotypes that we were unable to assign to a specific open reading frame (ORF; 73 genes with loss-of-function phenotypes—49 such genes assigned to an ORF) had the same probability of having a BLAST hit at a particular *P* value, and the same probability of having an EST match, as the 49 genes we could assign to single ORFs. We multiplied the number of the 49 genes with a phenotype that had a BLAST hit at a particular value of *P* or an EST match by 73/49 and then subtracted this number from the corresponding number derived using 218 genes in the *Adh* region.

[a] The percentage of these genes that also have EST matches is given in parentheses.

P values $<10^{-7}$ are unlikely to be false predictions. Yet in the total data set of 218 genes we see that the fraction that have ESTs increases the higher we set the expectation: for "all" species hits it is 48% at $P = 10^{-7}$, 53% at $P = 10^{-20}$, 60% at $P = 10^{-50}$, and 80% for $P = 10^{-100}$. Genes with mutant phenotypes have ESTs at an overall higher frequency than do those without phenotypes (Table 4). The observation that "conserved" genes are more highly expressed than are "nonconserved" genes, as judged by the occurrence of ESTs, was first made by Green *et al.* (1993) in their analysis of evolutionarily conserved regions in proteins. They suggested that highly expressed genes might be under a higher selection pressure. The similar bias in *C. elegans*, where genes with matches to proteins in distant taxa (*i.e.*, non-Nematodes) are three times more likely to have an EST than genes with no such match, was confirmed by an analysis of the *C. elegans* sequence (The *C. elegans* Sequencing Consortium 1998).

tRNA genes: An initial rush of enthusiasm mapped many tRNA genes by *in situ* hybridization to the polytene chromosomes and many of these were subsequently cloned and sequenced (*e.g.*, Kubli 1982). A total of 182 tRNA genes have so far been mapped in Drosophila (data from FlyBase), yet others remain to be discovered (*e.g.*, tryptophan and cysteine tRNAs). Many tRNA genes occur in clusters, either of isoaccepting or diverse tRNAs. A cluster of five glycine tRNAs was already known in the *Adh* region (Meng *et al.* 1988; 13 others are known). In addition we have identified a single glutamine tRNA (the first to be sequenced in Drosophila; *BG:DS01514.1*) and a single leucine tRNA (five others are known; *BG:DS03192.1*), four proline tRNAs (two others are known), one (*BG:DS04641.2*) immediately distal to the glycyl-tRNA cluster, and three (*BG:DS01486.2–.4*) just proximal to this cluster, immediately distal to *osp*. The 100-kb region between *noc* and *osp* therefore contains nine tRNA genes.

Transposable elements: About 12% of the genome of *D. melanogaster* is estimated to be composed of transposable element sequences, ribosomal DNA, and core histone genes (Laird and McCarthy 1968; Spradling and Rubin 1981). Seventeen elements have been recognized in the sequence of the *Adh* region; 6 are LINE-like elements (*G*, *F*, *Doc*, and *jockey*), 11 are retrotransposons with long terminal repeats (LTRs; *copia*, *roo*, *297*, *blood*, *mdg1*-like and *yoyo*; see Figures 1 and 2). This is an average spacing of 1 element per 171 kb. On the

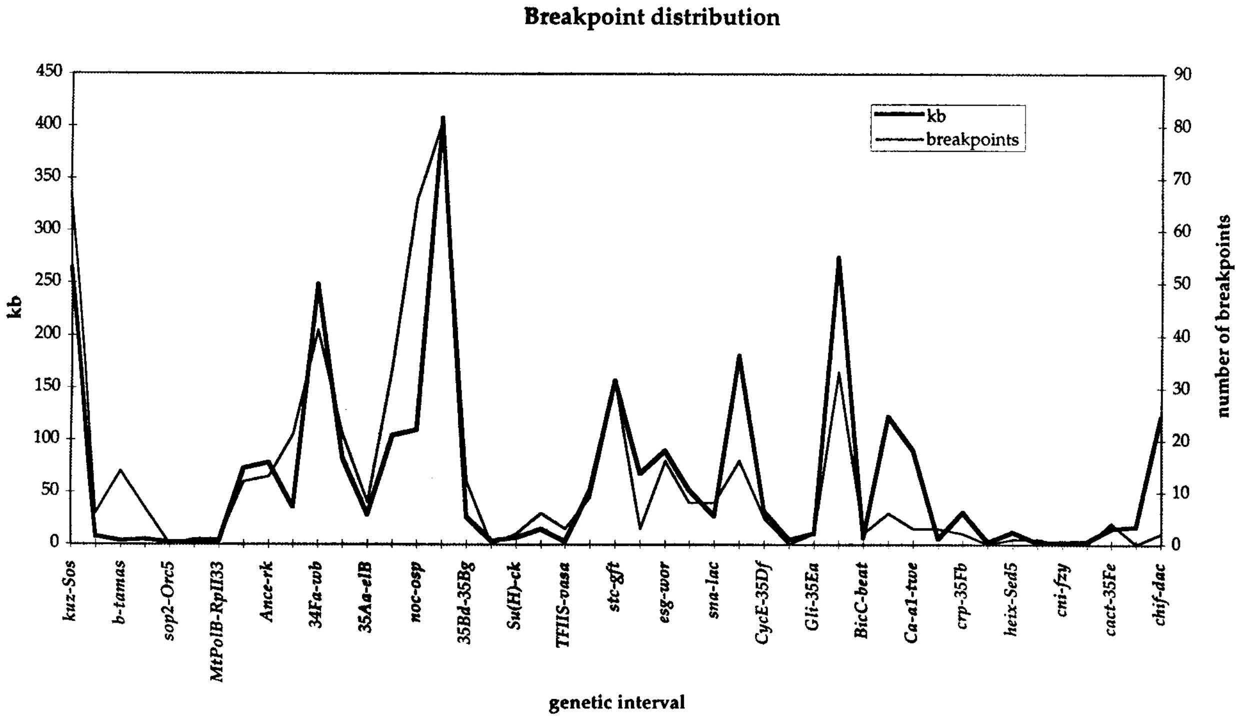


Figure 3.—A comparison of the distribution of DNA with that of genetically mapped chromosome breakpoints in the *kuz-dac* region. The genetic positions of 571 chromosome breakpoints have been determined (J. Roote, M. Ashburner and colleagues, data in FlyBase) with respect to 48 genes. The number in each gene interval is plotted along with the DNA content (in kilobases) of the same interval. The DNA lengths were measured between the chromosomally distal ends of genes (as defined by the predictions; see Figure 1).

basis of kinetic data the "middle-repetitive" sequences of *D. melanogaster* had been estimated to be ~5.6 kb in length, and separated by 13 kb or more of single-copy DNA (Manning *et al.* 1975; Crain *et al.* 1976).

A new retrotransposon element has been identified. It has been called *yoyo* in view of its sequence similarity with an element of the medfly *Ceratitis capitata* with this name. The *yoyo* LTR seems to be a hotspot for *P*-element insertion; *k08808*, a lethal allele of *l(2)35Bc*, is inserted in an LTR of *yoyo* and at least four other examples are known of *P* elements in *yoyo* LTRs (*PZ06264*, *EP(2)0533*, *EP(2)0396*, and *EP(2)0417*).

About 1.8% of the sequence of the *Adh* region is within identified transposable elements. This is much less than the 9% of the genome as a whole estimated to be composed of such sequences (Spradling and Rubin 1981). The reason for this difference is probably that the density of transposable elements is higher in the heterochromatic and peri-heterochromatic regions of the chromosomes (see Sun *et al.* 1997). Perhaps only half the retroviral elements are euchromatic. That this is so is indicated by a comparison of the total numbers of elements estimated by DNA reassociation kinetics and those seen in the euchromatic arms by *in situ* hybridization. For the *412* element, *e.g.*, the numbers were 40 (Potter *et al.* 1979) and 26 (Strobel *et al.* 1979), respectively, in Oregon-R; similar data were found for the *297* and *copia* elements.

There are other sequences that are clearly related to those of transposable elements but whose identity cannot be confidently stated. For example, on P1 clone DS07108 there are three very A + T-rich sequence regions that show similarities to elements such as *297* and *mdg1* but appear to be very degenerate. In addition, in an intron of *crp* there is an 860-bp sequence very similar to the repetitive element described as *Su(Ste)* (Balakireva *et al.* 1992).

Breakpoint distribution: We have mapped genetically 658 aberration breakpoints to this region of the Drosophila genome. Sixty-three breakpoints disrupt genes. Of these breakpoints many had previously been mapped to chromosome walks, usually in λ phage. Ninety-four of these were mapped to restriction fragments in the 450-kb "*Adh*" walk from Ashburner's laboratory (Chia *et al.* 1985; McGill *et al.* 1988; Davis *et al.* 1990, 1997; Gubb *et al.* 1990; Cheah *et al.* 1994; McNabb *et al.* 1996), while others had been mapped to the *vasa* (Lasko and Ashburner 1988), *Su(H)* (Schweisguth and Posakony 1992), *Sos* (Bonfini *et al.* 1992), *BicC* (Mahone *et al.* 1995), *beat* (Fambrough and Goodman 1996), *twe* (Alphey *et al.* 1992), *fzy* (Dawson *et al.* 1995), and *cni* (Roth *et al.* 1995) regions. Computer-generated restriction maps of the sequences of these regions were used to correlate these data with the sequence map. This was reasonably straightforward, the major adjustments being those needed to take transposable elements into account. We have compared the genetic and physical distributions of chromosome breakpoints in several ways. One is shown in Figure 3. In this figure we plot the numbers of breakpoints in each defined genetic interval with the length of DNA in that interval. It is clear that the two parameters are well correlated [Spearman's rank coefficient (Spearman 1904) $r_s = 0.78$, $t_{43} = 8.17$, $P = <0.001$], despite some degree of ascertainment bias in the data (most marked in the intervals surrounding *b* where very large-scale irradiation experiments have been done). Thus, the nonrandom clustering of aberration breakpoints seen in genetic mapping experiments is due to differing DNA target sizes rather than to some intrinsic property of the sequences themselves.

CONCLUSIONS

We chose the *Adh* region of *D. melanogaster* for our first experiment in megabase sequencing and sequence analysis because this region had been subjected to genetic analysis in greater detail than any of comparable size in a metazoan species. This has allowed us to integrate sequence analysis with saturating mutational analysis on a scale not previously seen in any metazoan organism.

A critical feature of the data is that the genes are not subject to ascertainment bias—they only share a common chromosomal location. The comparison of the sequences of genes known to be required for a "normal" phenotype and those not known by phenotypically mutant alleles has shown a surprisingly strong correlation between evolutionary conservation and "essentialness" of function. The fact that two independent measures of functional importance—evolutionary conservation over 500 million years and requirement for normal phenotype—are correlated has significant implications. For example, it argues that functionally essential genes are not organism specific, nor are their functions protected by gene duplication. Functionally essential genes show a second characteristic: on average they are expressed at higher levels, as judged by their representation in EST collections, than are genes that are not required for a normal phenotype.

Miklos and Rubin (1996) estimated that ~30% of the genes of *D. melanogaster* are "vital"; *i.e.*, loss of their function will result in lethality. Estimates of the fraction of the genes that are vital from our present analyses give the slightly lower figure of 24%, because we have 53 genes known or suspected from genetic data to be lethals, out of a total of 218 protein-coding genes.

One major challenge is to discover the functions of not only those genes for which mutant alleles are already known, but also those for which no alleles have been recovered in the screens performed so far. One general approach will be to engineer dominant gain-of-function alleles of these, *e.g.*, by using the *P* element engineered by Rørth (1996). Another approach will be to make double mutant combinations when we have reason to believe that a gene may be "redundant" due to a second gene in the genome. For example, mutations of *BG: DS08249.2* could be selected on a background mutant for the other known glycerol phosphate oxidase gene. Finally, the sequences or patterns of expression of a gene might suggest more appropriate phenotypic or biochemical assays to perform in search of its function.

This analysis of just 2.9 Mb of Drosophila sequence has been enormously informative and rewarding. Despite the fact that there is much more to be learned about this sequence, and the proteins it encodes, it has proved to be an invaluable experiment in preparation for the complete genomic sequence of this little fly, which we expect within the next year. Two matters are not in doubt; first, there is enough even in 2.9 Mb to keep biologists busy for many years and, second, their work will be invaluable in furthering our understanding not only of how Drosophila works and how it evolved, but also of human gene function.

We have benefited greatly from the resources of FlyBase, supported by grants from the National Institutes of Health (NIH) and the Medical Research Council (MRC, London). We thank K. Matthews and the Drosophila Stock Center, Bloomington, supported by grants from the National Science Foundation, both for keeping our own mutant strains and for supplying many others.

Many colleagues have made information available to us in advance of its publication. In particular we thank L. Alphey, M. Anderson, R. Anholt, J. Baker, U. Banerjee, S. Baumgartner, P. Benos, M. Botchan, S. D. M. Brown, A. Campos, D. Cavener, W. Chia, S. Cohen, I. Darboux, I. Dawson, J. B. Duffy, K. Edwards, G. Fedorowicz, C. Flores, J. Gates, M. Gatti, S. Hayashi, J. Heilig, A. Hudson, T. Ip, B. Iyengar, S. Jones, L. Keegan, D. Kiehart, I. Kiss, F. Laski, P. Lasko, M. Leptin, H. Mistry, T. Pipes, M. Pflumm, J. Posakony, J.-M. Reichhart, F. Schweisguth, K. Schmid, C. Sunkel, M. Taylor, C. Thummel, P. Tolias, J. Tower, N. Wakabayashi-Ito, A. Willingham, and last, but by no means least, C. Zuker.

We thank the following individuals who helped produce the sequence data presented in this article: A. Aghavani, D. A. Alcivare, T. T. Arcaina, D. Aragnol, E. Baxter, M. M. Bondoc, M. Chew, A. Chiang, P. A. Critz, I. Darboux, C. A. Davis, C. L. Ericsson, D. Fambrough, D. E. Farnan, J. Flanagan, K. M. Gunning, S. R. Hummasti, M. A. Jaklevic, K. E. Kadner, K. Karra, L. Kearney, K. Kim, S. F. Kim, S. H. Kim, C. L. Ko, B. Lee, K. D. Lewis, M. Li, K. J. Lindquist, M. Lomatan, V. M. Lustre, M. U. Machrus, C. A. Mayeda, P. Mazda, T. M. Miguel, C. A. Miller, M. S. Mok, M. Moshrefi, K. Nixon, J. M. Pacleb, S. Park, S. G. Patel, B. Pfeiffer, D. Punch, A. Salva, R. F. Santos, E. Snir, S. G. S. Subramanian, R. Svirskas, M. Taylor, B. Towne, B. Twomey, A. Yee, R. T. Yeh, C. Yu, R. Zhang, and L. L. Zieran. We thank S. Mullaney for drawing Figures 1 and 2.

Stocks or clones have been kindly provided to us by L. Alphey, C. Flores, J. Gates, L. Keegan, D. Kiehart, M. Phillips, T. Pipes, K. Tatei, P. Tolias and C. Zuker. We thank A. Brazma, European Bioinformatics Institute (EBI), for testing for randomness in the direction of transcription, W. Fleischmann (EBI) for the EMOTIF analyses, P. Horton (Osaka) for running the PSORT programs for us, J. Barrett (Cambridge) for statistical advice, T. Benos (EBI) for help with the sequence data sets, and R. Svirskas (Motorola Inc.) for providing informatics support at Berkeley. The help, advice, and programs of G. Helt (Berkeley) have been absolutely invaluable.

M.A. and G.M.R. thank both past and present members of their groups for both material and intellectual (not to say moral) support for this work: at Cambridge, S. Brogna, W. Chia, C. Detwiler, A. de

Grey, D. Gubb, D. Huen, R. Karp, D. Kimbrell, P. Lasko, S. McGill, S. McNabb, S. Tsubota, S. Russell, and R. Woodruff; at Berkeley, E. Frise, G. Mardon, D. J. Pan, M. Simon, and T. Xu. This work would have been quite impossible without the dedicated and skillful technical support of, at Cambridge, P. Thompson, D. Coulson, B. Durrant, J. Faithfull, P. Fletcher, S. Herrmann, T. Littlewood, T. Morley, M. Omar, M. Shelton, J. Trenear, and Y. Zhang; at Berkeley, A. Beaton, S. Chai, M. Evans-Holm, T. Laverty, D. Simas, and C. Suh.

G. M. Rubin thanks M. Bissell, A. Chatterjee, P. Oddone, C. Shank, and others at Lawrence Berkeley National Laboratory for their continuous support and encouragement, as well as L. Rubin for her patience. M.A., S.L., and S.M. thank W. M. Gelbart and his colleagues for their hospitality at Harvard.

Work in Cambridge on the genetic and molecular analyses of the *Adh* region has been continuously supported since 1983 by a MRC Programme Grant G8225539 to M. Ashburner and colleagues. Work in Berkeley was supported by the Drosophila Genome Center Grant from the NIH (P50 HG00750) and grants from the Department of Energy (DOE/DE-FG03-98ER62625 and DOE/DE-FG03-99ER62739) to G. Rubin and colleagues as well as by the Howard Hughes Medical Institute.

Note added in proof: Landis and Tower (1999) show that the *chiffon* protein shares two domains with Dbf4p of *S. cerevisiae*. *smi35A* has now been cloned and sequenced independently by Cleghon and colleagues (EMBL:AF168467); Min and Benzer (1999) have described *BG:DS0514.2* as *bubblegum* and have shown that a mutant allele has a neurodegeneration phenotype that can be corrected by feeding larvae glyceryl trioleate oil. Note that this gene is one of two in this region predicted to code for a long-chain fatty acid coenzyme A ligase (the other is *BG:DS05899.1*). *BG:DS00941.1* was identified as encoding a carbonate dehydratase by analysis of our sequence by Hewett-Emmett and Tashian (D. Hewett-Emmett and R. E. Tashian, 1996, Functional diversity, conservation, and convergence in the evolution of the α-, β-, and γ-carbonic anhydrase gene families. Mol. Phylogen. Evol. **5:** 50–77); they called this gene *CAH*, subsequently changed to *CAH1* (D. Hewett-Emmett, personal communication).

LITERATURE CITED

Achstetter, T., A. Franzusoff, C. Field and R. Schekman, 1988 SEC7 encodes an unusual, high molecular weight protein required for membrane traffic from the yeast Golgi apparatus. J. Biol. Chem. **263:** 11711–11717.

Adachi-Yamada, T., M. Nakamura, K. Irie, Y. Tomoyasu, Y. Sano *et al.*, 1999 p38 MAP kinase can be involved in TGF-beta superfamily signal transduction in *Drosophila* wing morphogenesis. Mol. Cell. Biol. **19:** 2322–2329.

Adams, C. M., M. G. Anderson, D. G. Motto, M. P. Price, W. A. Johnson *et al.*, 1998 Ripped pocket and pickpocket, novel *Drosophila* DEG/ENaC subunits expressed in early development and in mechanosensory neurons. J. Cell Biol. **140:** 143–152.

Alberga, A., J. L. Boulay, E. Kempe, C. Dennefeld and M. Haenlin, 1991 The snail gene required for mesoderm formation in *Drosophila* is expressed dynamically in derivatives of all three germ layers. Development **111:** 983–992.

Alphey, L., J. Jimenez, H. White-Cooper, I. Dawson, P. Nurse *et al.*, 1992 twine, a cdc25 homolog that functions in the male and female germline of *Drosophila.* Cell **69:** 977–988.

Anholt, R. R. H., R. F. Lyman and T. F. C. Mackay, 1996 Effects of single P-element insertions on olfactory behavior in *Drosophila melanogaster.* Genetics **143:** 293–301.

Ashburner, M., 1982 The genetics of a small autosomal region of *Drosophila melanogaster* containing the structural gene for Alcohol dehydrogenase. III. Hypomorphic and hypermorphic mutations affecting the expression of Hairless. Genetics **101:** 447–459.

Ashburner, M., 1989 *Drosophila: A Laboratory Handbook.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Ashburner, M., 1998 Speculations on the subject of alcohol dehydrogenase and its properties in *Drosophila* and other flies. Bioessays **20:** 949–954.

Ashburner, M., C. S. Aaron and S. Tsubota, 1982a The genetics of a small autosomal region of *D. melanogaster*, including the structural gene for Alcohol Dehydrogenase. V. Characterization of X-ray-induced *Adh* null mutations. Genetics **102:** 421–435.

Ashburner, M., S. Tsubota and R. C. Woodruff, 1982b The genetics of a small chromosome region of *Drosophila melanogaster* containing the structural gene for Alcohol dehydrogenase. IV. Scutoid, an antimorphic mutation. Genetics **102:** 401–420.

Ashburner, M., P. Thompson, J. Roote, P. F. Lasko, Y. Grau *et al.*, 1990 The genetics of a small autosomal region of *Drosophila melanogaster* containing the structural gene for alcohol dehydrogenase. VII. Characterization of the region around the snail and cactus loci. Genetics **126:** 679–694.

Auld, V. J., R. D. Fetter, K. Broadie and C. S. Goodman, 1995 Gliotactin, a novel transmembrane protein on peripheral glia, is required to form the blood-nerve barrier in *Drosophila.* Cell **81:** 757–767.

Baglioni, C., 1963 Correlations between genetics and chemistry of human haemoglobins, pp. 405–475 in *Progress in Molecular Genetics*, Vol. 1, edited by J. H. Taylor. Academic Press, New York.

Bahn, E., 1972 A suppressor locus for the pyrimidine requiring mutant: rudimentary. Dros. Inf. Serv. **49:** 98.

Bailey, A. M., and J. W. Posakony, 1995 Suppressor of Hairless directly activates transcription of Enhancer of split complex genes in response to Notch receptor activity. Genes Dev. **9:** 2609–2622.

Balakireva, M. D., Y. Y. Shevelyov, D. I. Nurminsky, K. J. Livak and V. A. Gvozdev, 1992 Structural organization and diversification of Y-linked sequences comprising *Su(Ste)* genes in *Drosophila melanogaster.* Nucleic Acids Res. **20:** 3731–3736.

Banfield, D. K., M. J. Lewis, C. Rabouille, G. Warren and H. R. B. Pelham, 1994 Localization of Sed5, a putative vesicle targeting molecule, to the cis-Golgi network involves both its transmembrane domain and cytoplasmic domains. J. Cell Biol. **127:** 357–371.

Barrett, J. A., 1980 The estimation of the number of mutationally silent loci in saturation-mapping experiments. Genet. Res. **35:** 33–44.

Barrett, A. J., N. D. Rawlings and J. F. Woessner, 1998 *Handbook of Proteolytic Enzymes.* Academic Press, San Diego.

Bass, B. L., 1997 RNA editing and hypermutation by adenosine deamination. Trends Biochem. Sci. **22:** 157–162.

Bateman, A., E. Birney, R. Durbin, S. R. Eddy, R. D. Finn *et al.*, 1999 Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res. **27:** 260–262.

Belote, J. M., F. M. Hoffmann, M. McKeown, R. L. Chorsky and B. S. Baker, 1990 Cytogenetic analysis of chromosome region 73AD of *Drosophila melanogaster.* Genetics **125:** 783–793.

Berkeley *Drosophila* Genome Project, 1999 http://www.fruitfly.org/.

Bevan, M., I. Bancroft, E. Bent, K. Love, H. Goodman *et al.*, 1998 Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana.* Nature **391:** 485–488.

Bomer, U., J. Rassow, N. Zufall, N. Pfanner, M. Meijer *et al.*, 1996 The preprotein translocase of the inner mitochondrial membrane: evolutionary conservation of targeting and assembly of Tim17. J. Mol. Biol. **262:** 389–395.

Bonfini, L., C. A. Karlovich, C. Dasgupta and U. Banerjee, 1992 The Son of sevenless gene product: a putative activator of Ras. Science **255:** 603–606.

Boulay, J. L., C. Dennefeld and A. Alberga, 1987 The *Drosophila* developmental gene snail encodes a protein with nucleic acid binding fingers. Nature **330:** 395–398.

Breitwieser, W., F. H. Markussen, H. Horstmann and A. Ephrussi, 1996 Oskar protein interaction with Vasa represents an essential step in polar granule assembly. Genes Dev. **10:** 2179–2188.

Brendel, V., P. Bucher, I. Nourbakhsh, B. E. Blaisdell and S. Karlin, 1992 Methods and algorithms for statistical analysis of protein sequences. Proc. Natl. Acad. Sci. USA **89:** 2002–2006.

Bridges, C. B., and K. S. Brehme, 1944 *The Mutants of* Drosophila melanogaster. Publs. Carnegie Instn. 552.

Brogna, S., and M. Ashburner, 1997 The Adh-related gene of *Drosophila melanogaster* is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms. EMBO J. **16:** 2023–2031.

Burge, C., 1997 Identification of genes in human genomic DNA. Ph.D. Thesis, Stanford University, Stanford, CA.

Burge, C., and S. Karlin, 1997 Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268:** 78–94.

C. elegans Sequencing Consortium, The, 1998 Genomic sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282:** 2012–2018.

C. elegans Genome Sequencing Project, The, 1999 How the worm was won. Trends Genet. **15:** 51–58.

Castle, L. A., and D. W. Meinke, 1994 A FUSCA gene of *Arabidopsis* encodes a novel protein essential for plant development. Plant Cell **6:** 25–41.

Castrillon, D. H., P. Gonczy, S. Alexander, R. Rawson, C. G. Eberhart *et al.*, 1993 Toward a molecular genetic analysis of spermatogenesis in *Drosophila melanogaster*: characterization of male-sterile mutants generated by single P-element mutagenesis. Genetics **135:** 489–505.

Cheah, P. Y., Y. B. Meng, X. Yang, D. A. Kimbrell, M. Ashburner *et al.*, 1994 The *Drosophila l(2)35Ba/nocA* gene encodes a putative Zn finger protein involved in the development of the embryonic brain and the adult ocellar structures. Mol. Cell. Biol. **14:** 1487–1499.

Chen, T. L., K. A. Edwards, R. C. Lin, L. W. Coats and D. P. Kiehart, 1991 *Drosophila* myosin heavy chain at 35BC. J. Cell Biol. **115:** 330a.

Chen, Z.-Y., T. Hasson, P. M. Kelley, B. J. Schwender, M. F. Schwartz *et al.*, 1996 Molecular cloning and domain structure of human myosin-VIIa, the gene product defective in Usher Syndrome 1B. Genomics **36:** 440–448.

Chia, W., R. Karp, S. McGill and M. Ashburner, 1985 Molecular analysis of the *Adh* region of the genome of *Drosophila melanogaster*. J. Mol. Biol. **186:** 689–706.

Chiu, S. K., and M. A. Krasnow, 1997 Identification of new genes required for the formation of terminal tracheal branches. A. Conf. Dros. Res. **38:** 229A.

Choudhary, M., M. B. Coulthart and R. S. Singh, 1992 A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. VI. Patterns and processes of genic divergence between *Drosophila melanogaster* and its sibling species, *Drosophila simulans*. Genetics **130:** 843–853.

Cornell, M. J., T. A. Williams, N. S. Lamango, D. Coates, P. Corvol *et al.*, 1995 Cloning and expression of an evolutionary conserved single-domain angiotensin converting enzyme from *Drosophila melanogaster*. J. Biol. Chem. **270:** 13613–13619.

Courtot, C., C. Fankhauser, V. Simanis and C. F. Lehner, 1992 The *Drosophila cdc25* homolog twine is required for meiosis. Development **116:** 405–416.

Crain, W. R., F. C. Eden, W. R. Pearson, E. H. Davidson and R. J. Britten, 1976 Absence of short period interspersion of repetitive and non-repetitive sequences in the DNA of *Drosophila melanogaster*. Chromosoma **56:** 309–326.

Critchlow, S. E., and S. P. Jackson, 1998 DNA end-joining: from yeast to man. Trends Biochem. Sci. **23:** 394–398.

Cutler, M. L., R. H. Bassin, L. Zanoni and N. Talbot, 1992 Isolation of rsp-1, a novel cDNA capable of suppressing v-Ras transformation. Mol. Cell. Biol. **12:** 3750–3756.

Danielson, P. B., R. J. MacIntyre and J. C. Fogleman, 1997 Molecular cloning of a family of xenobiotic-inducible drosophilid cytochrome p450s: evidence for involvement in host-plant allelochemical resistance. Proc. Natl. Acad. Sci. USA **94:** 10797–10802.

Darboux, I., E. Lingueglia, D. Pauron, P. Barbry and M. Lazdunski, 1998 A new member of the amiloride-sensitive sodium channel family in *Drosophila melanogaster* peripheral nervous system. Biochem. Biophys. Res. Commun. **246:** 210–216.

Davis, M. B., and R. J. MacIntyre, 1988 A genetic analysis of the α-glycerophosphate oxidase locus in *Drosophila melanogaster*. Genetics **120:** 755–766.

Davis, T., J. Trenear and M. Ashburner, 1990 The molecular analysis of the *el-noc* complex of *Drosophila melanogaster*. Genetics **126:** 105–119.

Davis, T., M. Ashburner, G. Johnson, D. Gubb and J. Roote, 1997 Genetic and phenotypic analysis of the genes of the elbow-no-ocelli region of chromosome 2L of *Drosophila melanogaster*. Hereditas **126:** 67–75.

Dawson, I. A., S. Roth, M. Akam and S. Artavanis-Tsakonas, 1993 Mutations of the fizzy locus cause metaphase arrest in *Drosophila melanogaster* embryos. Development **117:** 359–376.

Dawson, I. A., S. Roth and S. Artavanis-Tsakonas, 1995 The *Drosophila* cell cycle gene fizzy is required for normal degradation of cyclins A and B during mitosis and has homology to the CDC20 gene of *Saccharomyces cerevisiae*. J. Cell Biol. **129:** 725–737.

De La Vega, H., C. A. Specht, Y. Liu and P. W. Robbins, 1998 Chitinases are a multi-gene family in *Aedes*, *Anopheles* and *Drosophila*. Insect Mol. Biol. **7:** 233–239.

De Vries, L., M. Mousli, A. Wurmser and M. G. Farquhar, 1995 GAIP, a protein that specifically interacts with the trimeric G protein G alpha i3, is a member of a protein family with a highly conserved core domain. Proc. Natl. Acad. Sci. USA **92:** 11916–11920.

Eberl, D. F., D. Ren, G. Feng, L. J. Lorenz, D. Van Vactor *et al.*, 1998 Genetic and developmental characterization of *Dmca1D*, a calcium channel α1 subunit gene in *Drosophila melanogaster*. Genetics **148:** 1159–1169.

Eddy, S. R., 1998 HAMMER2.1 Profile hidden Markov models for biological sequence analysis. http://hmmer.wustl.edu/.

Edgar, B. A., 1994 Cell cycle. Cell-cycle control in a developmental context. Curr. Biol. **4:** 522–524.

Edmondson, M. E., 1948 New mutants report. Dros. Inf. Serv. **22:** 53.

European *Drosophila* Genome Project, 1999 http://edgp.ebi.ac.uk/.

Fambrough, D., and C. S. Goodman, 1996 The *Drosophila* beaten path gene encodes a novel secreted protein that regulates defasciculation at motor axon choice points. Cell **87:** 1049–1058.

Fambrough, D., D. Pan, G. M. Rubin and C. S. Goodman, 1996 The cell surface metalloprotease/disintegrin Kuzbanian is required for axonal extension in *Drosophila*. Proc. Natl. Acad. Sci. USA **93:** 13233–13238.

Florea, L., G. Hartzell, Z. Zhang, G. M. Rubin and W. Miller, 1998 A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. **8:** 967–974.

Flores, C., and W. R. Engels, 1999 Microsatellite instability in *Drosophila spellchecker1* (MutS homolog) mutants. Proc. Natl. Acad. Sci. USA **96:** 2964–2969.

FlyBase Consortium, 1999 The FlyBase database of the *Drosophila* genome projects and community literature. Nucleic Acids Res. **27:** 85–88.

Frank, L. H., and C. Rushlow, 1996 A group of genes required for maintenance of the amnioserosa tissue in *Drosophila*. Development **122:** 1343–1352.

Franzusoff, A., K. Redding, J. Crosby, R. S. Fuller and R. Schekman, 1991 Localization of components involved in protein transport and processing through the yeast Golgi apparatus. J. Cell Biol. **112:** 27–37.

Fuchs, R., 1994 Predicting protein functions: a versatile tool for the Apple Macintosh. CABIOS **10:** 171–178.

Furukawa, T., S. Maruyama, M. Kawaichi and T. Honjo, 1992 The *Drosophila* homolog of the immunoglobulin recombination signal-binding protein regulates peripheral nervous system development. Cell **69:** 1191–1197.

Fuse, N., S. Hirose and S. Hayashi, 1996 Determination of wing cell fate by the escargot and snail genes in *Drosophila*. Development **122:** 1059–1067.

Gausz, J., G. Bencze, H. Gyurkovics, M. Ashburner, D. Ish-Horowicz *et al.*, 1979 Genetic characterization of the 87C region of the third chromosome of *Drosophila melanogaster*. Genetics **93:** 917–934.

Geisler, R., A. Bergmann, Y. Hiromi and C. Nusslein-Volhard, 1992 cactus, a gene involved in dorsoventral pattern formation of *Drosophila*, is related to the IκB gene family of vertebrates. Cell **71:** 613–621.

Gibson, F., J. Walsh, P. Mburu, A. Varela, K. A. Brown *et al.*, 1995 A type VII myosin encoded by the mouse deafness gene *shakyer-1*. Nature **374:** 62–64.

Gene Ontology Consortium, 1999 http://www.ebi.ac.uk/~ashburn/GO/ and http://www.fruitfly.org/~suzi/.

Gonzalez-Reyes, A., H. Elliott and R. D. St. Johnston, 1995 Polarization of both major body axes in *Drosophila* by gurken-torpedo signalling. Nature **375:** 654–658.

Gossen, M., D. T. S. Pak, S. K. Hansen, J. K. Acharya and M. R. Botchan, 1995 A *Drosophila* homolog of the yeast origin recognition complex. Science **270:** 1674–1677.

Grau, V., G. Carteret and P. Simpson, 1984 Mutation and chromosomal rearrangements affecting the expression of snail, a gene involved in embryonic patterning in *Drosophila melanogaster*. Genetics **108:** 347–360.

Green, E. D., and M. V. Olson, 1990 Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. Proc. Natl. Acad. Sci. USA **87:** 1213–1217.

Green, P., 1995 GENEFINDER Documentation. http://www.ibc.wustl.edu/bio_data/genefinder.html.

Green, P., D. Lipman, L. Hillier, R. Waterston, D. States *et al.*, 1993 Ancient conserved regions in new gene sequences and the protein databases. Science **259:** 1711–1716.

Grell, E. H., K. B. Jacobson and J. B. Murphy, 1968 Alterations of genetic material for analysis of alcohol dehydrogenase isozymes of *Drosophila melanogaster*. Ann. NY Acad. Sci. **151:** 441–455.

Griffith, J. K., and C. E. Sansom, 1998 *The Transporter Facts Book*. Academic Press, San Diego.

Gubb, D., 1998 Chromosome mechanics: the genetic manipulation of aneuploid stocks, pp. 109–130 in *Drosophila: A Practical Approach*, edited by D. B. Roberts. IRL Press, Oxford.

Gubb, D., M. Shelton, J. Roote, S. McGill and M. Ashburner, 1984 The genetic analysis of a large transposing element of *Drosophila melanogaster*. The insertion of a w^+ rst^+ TE into the *ck* locus. Chromosoma **91:** 54–64.

Gubb, D., J. Roote, G. Harrington, S. McGill, B. Durrant *et al.*, 1985 A preliminary genetic analysis of *TE146*, a very large transposing element of *Drosophila melanogaster*. Chromosoma **92:** 116–123.

Gubb, D., M. Ashburner, J. Roote and T. Davis, 1990 A novel transvection phenomenon affecting the *white* gene of *Drosophila melanogaster*. Genetics **126:** 167–176.

Guo, M., L. Y. Jan and Y. N. Jan, 1996 Control of daughter cell fates during asymmetric division: interaction of *numb* and *Notch*. Neuron **17:** 27–41.

Han, Z. S., H. Enslen, X. Hu, X. Meng, I.-H. Wu *et al.*, 1998 A conserved p38 mitogen-activated protein kinase pathway regulates *Drosophila* immunity gene expression. Mol. Cell. Biol. **18:** 3527–3539.

Hartl, D. L., D. I. Nurminsky, R. W. Jones and E. R. Lozovskaya, 1994 Genome structure and evolution in *Drosophila*: applications of the framework P1 map. Proc. Natl. Acad. Sci. USA **91:** 6824–6829.

Hauser, F., H. P. Nothacker and C. J. Grimmelikhuijzen, 1997 Molecular cloning, genomic organization, and developmental regulation of a novel receptor from *Drosophila melanogaster* structurally related to members of the thyroid-stimulating hormone, follicle-stimulating hormone, luteinizing hormone/choriogonadotropin receptor family from mammals. J. Biol. Chem. **272:** 1002–1010.

Hay, B. A., L. Y. Jan and Y. N. Jan, 1988 A protein component of *Drosophila* polar granules is encoded by vasa and has extensive sequence similarity to ATP-dependent helicases. Cell **55:** 577–587.

Hayashi, S., 1996 Checkpoint mechanism that maintains diploidy in *Drosophila*: CDC2 inhibits S phase entry in G2 by a kinase independent mechanism. Cell Struct. Funct. **21:** 694.

Hayashi, S., S. Hirose, T. Metcalfe and A. D. Shirras, 1993 Control of imaginal cell development by the escargot gene of *Drosophila*. Development **118:** 105–115.

Heitzler, P., D. Coulson, M. T. Saenz-Robles, M. Ashburner, J. Roote *et al.*, 1993 Genetic and cytogenetic analysis of the 43A-E region containing the segment polarity gene costa and the cellular polarity genes prickle and spiny-legs in *Drosophila melanogaster*. Genetics **135:** 105–115.

Helt, G., 1997 Data visualization and gene discovery in *Drosophila melanogaster*. Ph.D. Thesis, University of California, Berkeley, CA.

Henikoff, S., M. A. Keene, K. Fechtel and J. W. Fristrom, 1986 Gene within a gene: nested *Drosophila* genes encode unrelated proteins on opposite DNA strands. Cell **44:** 33–42.

Higgins, D. G., J. D. Thompson and T. J. Gibson, 1996 Using CLUSTAL for multiple sequence alignments. Methods Enzymol. **266:** 383–402.

Hilliker, A. J., S. H. Clark, W. M. Gelbart and A. Chovnick, 1981 Cytogenetic analysis of the rosy micro-region, polytene chromosome interval 87D2-4; 87E12-F1, of *D. melanogaster*. Dros. Inf. Serv. **56:** 65–72.

Hodgetts, R. B., 1972 Biochemical characterization of mutants affecting the metabolism of β-alanine in *Drosophila*. J. Insect Physiol. **18:** 937–947.

Hofmann, K., P. Bucher, L. Falquet and A. Bairoch, 1999 The PROSITE database, its status in 1999. Nucleic Acids Res. **27:** 215–219.

Holmes, A. L., and J. S. Heilig, 1998 Fascilin II and beaten path modulate intercellular adhesion in larval visual organ development. Development **126:** 261–272.

Holmes, A. L., R. N. Raper and J. S. Heilig, 1998 Genetic analysis of *Drosophila* larval optic nerve development. Genetics **148:** 1189–1201.

Horton, P., and K. Nakai, 1997 Better prediction of protein cellular localization sites with the k nearest neighbors classifier. Proc. Int. Conf. Intelligent Syst. Mol. Biol. **5:** 147–152.

Hosie, A. M., K. Aronstein, D. B. Sattelle and R.H. ffrench-Constant, 1997 Molecular biology of insect neuronal GABA receptors. Trends Neurosci. **20:** 578–583.

Houard, X., T. A. Williams, A. Michaud, D. Dani, R. E. Isaac *et al.*, 1998 The *Drosophila melanogaster*-related angiotensin-I-converting enzymes Acer and Ance. Distinct enzymic characteristics and alternative expression during pupal development. Eur. J. Biochem. **257:** 599–606.

Hudson, A., and L. Cooley, 1998 Analysis of the *Drosophila* Arp2/3 complex in oogenesis. A. Dros. Res. Conf. **39:** 289B.

Hwang, S.-Y., B. Oh, Z. Zhang, W. Miller, D. Solter *et al.*, 1999 The mouse *cornichon* gene family. Dev. Genes Evol. **209:** 120–125.

Ingram, V. N., 1961 Gene evolution and the haemoglobins. Nature **189:** 704–708.

Iwaki, D., S. Kawabata, Y. Miura, A. Kato, P. B. Armstrong *et al.*, 1996 Molecular cloning of Limulus alpha 2-macroglobulin. Eur. J. Biochem. **242:** 822–831.

Jackson, F. R., L. M. Newby and S. J. Kulkarni, 1990 Drosophila GABAergic systems: sequence and expression of glutamic acid decarboxylase. J. Neurochem. **54:** 1068–1078.

Jacobs, M. E., 1974 Beta-alanine and adaptation in *Drosophila*. J. Insect Physiol. **20:** 859–866.

Jimenez, J., L. Alphey, P. Nurse and D. M. Glover, 1990 Complementation of fission yeast *cdc2ts* and *cdc25ts* mutants identifies two cell cycle genes from *Drosophila*: a *cdc2* homologue and string. EMBO J. **9:** 3565–3571.

Jones, S. J. M., 1999 Computational analysis of the *Caenorhabditis elegans* genome sequence. Ph.D. Thesis, Open University, England.

Judd, B. H., M. W. Shen and T. C. Kaufman, 1972 The anatomy and function of a segment of the X chromosome of *Drosophila melanogaster*. Genetics **71:** 139–156.

Kamizono, A., M. Nishizawa, Y. Teranishi, K. Murata and A. Kimura, 1989 Identification of a gene conferring resistance to zinc and cadmium ions in the yeast *Saccharomyces cerevisiae*. Mol. Gen. Genet. **219:** 161–167.

Karlstrom, R. O., L. P. Wilder and M. J. Bastiani, 1993 Lachesin: an immunoglobulin superfamily protein whose expression correlates with neurogenesis in grasshopper embryos. Development **118:** 509–522.

Kavenoff, R., and B. H. Zimm, 1973 Chromosome-sized DNA molecules from *Drosophila*. Chromosoma **41:** 1–27.

Kawabata, S., F. Tokunaga, Y. Kugi, S. Motoyama, Y. Miura *et al.*, 1996 *Limulus* factor D, a 43-kDa protein isolated from horseshoe crab hemocytes, is a serine protease homologue with antimicrobial activity. FEBS Lett. **398:** 146–150.

Kawamura, K., T. Shibata, O. Saget, D. Peel and P. J. Bryant, 1999 A new family of growth factors produced by the fat body and active on *Drosophila* imaginal disc cells. Development **126:** 211–219.

Kimmel, B. E., M. J. Palazzolo, C. H. Martin, J. D. Boeke and S. E. Devine, 1997 Transposon-mediated DNA sequencing, pp. 455–532 in *Genome Analysis*, Vol. 1, edited by B. Birren, E. D. Green, S. Klapholz, R. M. Myers and J. Roskams. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Kimmerly, W. J., K. Stultz, S. Lewis, K. Lewis, V. Lustre *et al.*, 1996 A P1-based physical map of the *Drosophila* euchromatic genome. Genome Res. **6:** 414–430.

Kobayashi, S., S. Miyabe, S. Izawa, Y. Inoue and A. Kimura, 1996 Correlation of the OSR/ZRCI gene product and the intracellular glutathione levels in *Saccharomyces cerevisiae*. Biotechnol. Appl. Biochem. **23:** 3–6.

Kohler, R. E., 1994 *Lords of the Fly: Drosophila Genetics and the Experimental Life*. University of Chicago Press, Chicago.

Kozlova, T. Y., V. F. Semeshin, I. V. Tretyakova, E. B. Kokoza, V. Pirrotta *et al.*, 1994 Molecular and cytogenetical characterization of the 10A1-2 band and adjoining region in the *Drosophila melanogaster* polytene X chromosome. Genetics **136:** 1063–1073.

Kramer, K. M., D. Fesquet, A. L. Johnson and L. H. Johnston, 1998 Budding yeast RSI1 / APC2, a novel gene necessary for initiation of anaphase, encodes an APC subunit. EMBO J. **17:** 498–506.

Kubli, E., 1982 The genetics of transfer RNA in *Drosophila*. Adv. Genet. **21:** 123–172.

Laird, C. D., 1971 Chromatid structure: relationship between DNA content and nucleotide sequence diversity. Chromosoma **32:** 378–406.

Laird, C. D., and B. J. McCarthy, 1968 Nucleotide sequence homology within the genome of *Drosophila melanogaster*. Genetics **60:** 323–334.

Laird, C. D., and B. J. McCarthy, 1969 Molecular characterization of the *Drosophila* genome. Genetics **63:** 865–882.

Lammer, D., N. Mathias, J. M. Laplaza, W. Jiang, Y. Liu *et al.*, 1998 Modification of yeast Cdc53p by the ubiquitin-related protein rub1p affects function of the SCF^{Cdc4} complex. Genes Dev. **12:** 914–926.

Landis, G., and J. Tower, 1999 The *Drosophila chiffon* gene is required for chorion gene amplification, and is related to the yeast Dbf4 regulator of DNA replication and cell cycle. Development **126** (in press).

Lasko, P. F., and M. Ashburner, 1988 The product of the *Drosophila* gene *vasa* is very similar to eukaryotic initiation factor 4A. Nature **335:** 611–617.

Lasko, P. F., and M. Ashburner, 1990 Posterior localization of vasa protein correlates with, but is not sufficient for, pole cell development. Genes Dev. **4:** 905–921.

Lee, E. C., S. Y. Yu, X. Hu, M. Mlodzik and N. E. Baker, 1998 Functional analysis of the fibrinogen-related scabrous gene from *Drosophila melanogaster* identifies potential effector and stimulatory protein domains. Genetics **150:** 663–673.

Lefevre, G., 1976 A photographic representation and interpretation of the polytene chromosomes of *Drosophila melanogaster* salivary glands, pp. 31–66 in *The Genetics and Biology of Drosophila*, Vol. 1a, edited by M. Ashburner and E. Novitski. Academic Press, London.

Lefevre, G., and W. S. Watkins, 1986 The question of the total gene number in *Drosophila melanogaster*. Genetics **113:** 869–895.

Leptin, M., 1994 Morphogenesis: control of epithelial cell shape changes. Curr. Biol. **4:** 709–712.

Lewis, E. B., J. D. Knafels, D. R. Mathog and S. E. Celniker, 1995 Sequence analysis of the cis-regulatory regions of the bithorax complex of *Drosophila*. Proc. Natl. Acad. Sci. USA **92:** 8403–8407.

Lewis, D. L., C. L. Farr, Y. Wang, A. T. Lagina and L. S. Kaguni, 1996 Catalytic subunit of mitochondrial DNA polymerase from *Drosophila* embryos: cloning, bacterial overexpression, and biochemical characterization. J. Biol. Chem. **271:** 23389–23394.

Lim, R., and A. Zaheer, 1996 In vitro enhancement of p38 mitogen-activated protein kinase activity by phosphorylated glia maturation factor. J. Biol. Chem. **271:** 22953–22956.

Lindsley, D. L., and G. G. Zimm, 1992 *The Genome of Drosophila melanogaster*. Academic Press, San Diego.

Littleton, J. T., and H. J. Bellen, 1994 Genetic and phenotypic analysis of thirteen essential genes in cytological interval 22F1-2; 23B1-2 reveals novel genes required for neural development in *Drosophila*. Genetics **138:** 111–123.

Lohe, A. R., and D. L. Brutlag, 1987 Adjacent satellite DNA segments in *Drosophila*. J. Mol. Biol. **194:** 171–179.

Lowe, T. M., and S. R. Eddy, 1997 tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequences. Nucleic Acids Res. **25:** 955–964.

Mahone, M., E. E. Saffman and P. F. Lasko, 1995 Localized Bicaudal-C RNA encodes a protein containing a KH domain, the RNA binding motif of FMR1. EMBO J. **14:** 2043–2055.

Maleszka, R., H. G. de Couet and G. L. G. Miklos, 1998 Data transferability from model organisms to human beings: insights from the functional genomics of the flightless region of *Drosophila*. Proc. Natl. Acad. Sci. USA **95:** 3731–3736.

Manning, J. E., C. W. Schmid and N. Davidson, 1975 Interspersion of repetitive and nonrepetitive DNA sequences in the *Drosophila melanogaster* genome. Cell **4:** 141–155.

Mardon, G., N. M. Solomon and G. M. Rubin, 1994 dachshund encodes a nuclear protein required for normal eye and leg development in *Drosophila*. Development **120:** 3473–3486.

Marrs, J. A., and G. B. Bouck, 1992 The two major membrane skeletal proteins (articulins) of *Euglena gracilis* define a novel class of cytoskeletal proteins. J. Cell Biol. **118:** 1465–1475.

Marshall, T. K., H. Guo and D. H. Price, 1990 Drosophila RNA polymerase II elongation factor DmS-II has homology to mouse S-II and sequence similarity to yeast PPR2. Nucleic Acids Res. **18:** 6293–6298.

Martin, C. H., C. A. Mayeda, C. A. Davis, C. L. Ericsson, J. D. Knafels *et al.*, 1995 Complete sequence of the bithorax complex of *Drosophila*. Proc. Natl. Acad. Sci. USA **92:** 8398–8402.

Martin, D., S. Zusman, X. Li, E. L. Williams, N. Khare *et al.*, 1999 *wing blister*, a new *Drosophila* laminin α chain required for cell adhesion and migration during embryonic and imaginal development. J. Cell Biol. **145:** 191–201.

McGill, S., 1985 Molecular studies of the *Adh* region of *Drosophila melanogaster*. Ph.D. Thesis, University of Cambridge, England.

McGill, S., W. Chia, R. Karp and M. Ashburner, 1988 The molecular analysis of an antimorphic mutation of *Drosophila melanogaster*, Scutoid. Genetics **119:** 647–661.

McKray, R. D., L. Zhu and R. D. Shortridge, 1995 A *Drosophila* gene that encodes a member of the protein disulfide isomerase / phospholipase C-α family. Insect Biochem. Mol. Biol. **25:** 647–654.

McNabb, S., S. Greig and T. Davis, 1996 The alcohol dehydrogenase gene is nested in the outspread locus of *Drosophila melanogaster*. Genetics **143:** 897–911.

Mello, C. C., B. W. Draper and J. R. Priess, 1994 The maternal genes *apx-1* and *glp-1* and establishment of dorsal-ventral polarity in the early *C. elegans* embryo. Cell **77:** 95–106.

Meng, Y. B., R. D. Stevens, W. Chia, S. McGill and M. Ashburner, 1988 Five glycyl tRNA genes within the noc gene complex of *Drosophila melanogaster*. Nucleic Acids Res. **16:** 7189.

Mewes, H. W., K. Albermann, M. Bähr, D. Frishman, A. Gliessner *et al.*, 1997 Overview of the yeast genome. Nature **387** (Suppl.): 7–8.

Miklow, G. L. G., and G. M. Rubin, 1996 The role of the genome project in determining gene function: insights from model organisms. Cell **86:** 521–529.

Milne, A. A., 1926 *Winnie-the-Pooh*. Methuen, London.

Min, K.-T., and S. Benzer, 1999 Preventing neurodegeneration in the *Drosophila* mutant bubblegum. Science **284:** 1985–1988.

Mistry, H., 1997 Identification of loci interacting with $G\alpha_s$ signalling in *Drosophila melanogaster*. Ph.D. Thesis, University of Cambridge, England.

Mohler, J., and E. Wieschaus, 1986 Dominant maternal-effect mutations of *Drosophila melanogaster* causing the production of double-abdomen embryos. Genetics **112:** 803–822.

Moringa, N., S. C. Tsai, J. Moss and J. Vaughan, 1996 Isolation of a brefeldin A-inhibited guanine nucleotide-exchange protein for ADP ribosylation factor (ARF) 1 and ARF3 that contains a Sec7-like domain. Proc. Natl. Acad. Sci. USA **93:** 12856–12860.

Munro, S., and H. R. Pelham, 1987 A C-terminal signal prevents secretion of luminal ER proteins. Cell **48:** 899–907.

Munroe, D. J., R. Loebbert, E. Bric, T. Whitton, D. Prawitt *et al.*, 1995 Systematic screening of an arrayed cDNA library by PCR. Proc. Natl. Acad. Sci. USA **92:** 2209–2213.

Murphy, S. M., L. Urbani and T. Stearns, 1998 The mammalian gamma-tubulin complex contains homologues of the yeast spindle pole body components spc97p and spc98p. J. Cell Biol. **141:** 663–674.

Musacchio, M., and N. Perrimon, 1996 The *Drosophila* kekkon genes: novel members of both the leucine-rich repeat and immunoglobulin superfamilies expressed in the CNS. Dev. Biol. **178:** 63–76.

Nakai, M., T. Endo, T. Hase and H. Matsubara, 1993 Intramitochondrial protein sorting: isolation and characterization of the yeast MSP1 gene which belongs to a novel family of putative ATPases. J. Biol. Chem. **268:** 24262–24269.

Nash, D., 1965 The expression of 'Hairless' in *Drosophila* and the role of two closely linked modifiers of opposite effect. Genet. Res. **6:** 175–189.

Neer, E. J., C. J. Schmidt, R. Nambudripad and T. F. Smith, 1994 The ancient regulatory-protein family of WD-repeat proteins. Nature **371:** 297–300.

Nevill-Manning, C. G., T. D. Wu and D. L. Brutlag, 1998 Highly specific protein sequence motifs for genome analysis. Proc. Natl. Acad. Sci. USA **95:** 5865–5871.

Norrander, J. M., A. Perrone, L. A. Amos and R. W. Linck, 1996 Structural comparison of tektins and evidence for their determination of complex spacings in flagellar microtubules. J. Mol. Biol. **257:** 385–397.

Nusslein-Volhard, D., E. Wieschaus and G. Jurgens, 1982 Segmentierung bei *Drosophila*. Verh. Ges. Dtsch. Zool. **1982:** 91–104.

Nusslein-Volhard, C., E. Wieschaus and H. Kluding, 1984 Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*. I. Zygotic loci on the second chromosome. Roux's Arch. Dev. Biol. **193:** 267–282.

O'Donnell, J. M., H. C. Mandel, M. Krauss and W. Sofer, 1977 Genetic and cytogenetic analysis of the *Adh* region in *Drosophila melanogaster*. Genetics **86:** 553–566.

Oh, Y., J. Yoon and K. Baek, 1995 Isolation and characterization of the gene encoding the *Drosophila melanogaster* transcriptional elongation factor, TFIIS. Biochim. Biophys. Acta **1262:** 99–103.

Olson, M. V., L. Hood, C. Cantor and D. Bostein, 1989 A common language for physical mapping of the human genome. Science **245:** 1434–1435.

Oppenheimer, D. G., M. A. Pollock, J. Vacik, D. B. Szymanski, B. Ericson *et al.*, 1997 Essential role of a kinesin-like protein in *Arabidopsis* trichome morphogenesis. Proc. Natl. Acad. Sci. USA **94:** 6261–6266.

Osoegawa, K., P. Y. Woon, B. Zhao, E. Frengen, M. Tateno *et al.*, 1998 An improved approach for construction of bacterial artificial chromosome libraries. Genomics **52:** 1–8.

Pan, D., and G. M. Rubin, 1997 Kuzbanian controls proteolytic processing of Notch and mediates lateral inhibition during *Drosophila* and vertebrate neurogenesis. Cell **90:** 271–280.

Patel, S., and M. Latterich, 1998 The AAA team: related ATPases with diverse functions. Trends Cell Biol. **8:** 65–71.

Pawson, T., and J. D. Scott, 1997 Signalling through scaffold, anchoring, adaptor proteins. Science **278:** 2075–2080.

Pedersen, M. B., 1982 Enhancement and suppression of the black mutant and induction of black phenocopies in *Drosophila melanogaster*. Hereditas **97:** 329.

Phillips, A. M., L. B. Salkoff and L. E. Kelly, 1993 A neural gene from *Drosophila melanogaster* with homology to vertebrate and invertebrate glutamate decarboxylases. J. Neurochem. **61:** 1291–1301.

Pieri, A., F. Magherini, G. Liguri, G. Raugei, N. Taddei *et al.*, 1998 *Drosophila melanogaster* acylphosphatase: a common ancestor for acylphosphatase isoenzymes of vertebrate species. FEBS Lett. **433:** 205–210.

Pinter, M., G. Jekely, R. J. Szepsesi, A. Farkas, U. Theopold *et al.*, 1998 TER94, a Drosophila homolog of the membrane fusion protein CDC48/p97, is accumulated in nonproliferating cells: in the reproductive organs and in the brain of the imago. Insect Biochem. Mol. Biol. **28:** 91–98.

Porter, T. G., and D. L. Martin, 1988 Non-steady state kinetics of brain glutamate decarboxylase resulting from the interconversion of the apo- and holoenzyme. Biochim. Biophys. Acta **874:** 235–244.

Potter, S. S., W. J. Brorein, P. Dunsmuir and G. M. Rubin, 1979 Transcription of elements of the *412*, *copia* and *297* dispersed repeated gene families in *Drosophila*. Cell **17:** 415–427.

Powers, J., and C. Barlowe, 1998 Transport of Ax12p depends on the Erv14p, an ER-vesicle protein related to the *Drosophila cornichon* gene product. J. Cell Biol. **142:** 1209–1222.

Rasch, E. M., H. J. Barr and R. W. Rasch, 1971 The DNA content of sperm of *Drosophila melanogaster*. Chromosoma **33:** 1–18.

Reese, M. G., F. H. Eeckman, D. Kulp and D. Haussler, 1997 Improved splice site detection in Genie. J. Comput. Biol. **4:** 311–323.

Richardson, H. E., L. V. O'Keefe, S. I. Reed and R. Saint, 1993 A *Drosophila* G1-specific cyclin E homolog exhibits different modes of expression during embryogenesis. Development **119:** 673–690.

Rogge, R. D., C. A. Karlovich and U. Banerjee, 1991 Genetic dissection of a neurodevelopmental pathway: son of sevenless functions downstream of the sevenless and EGF receptor tyrosine kinases. Cell **64:** 39–48.

Rooke, J., D. Pan, T. Xu and G. M. Rubin, 1996 KUZ, a conserved metalloprotease-disintegrin protein with two roles in *Drosophila* neurogenesis. Science **273:** 1227–1231.

Ropp, P. A., and W. C. Copeland, 1996 Cloning and characterization of the human mitochondrial DNA polymerase, DNA polymerase α. Genomics **36:** 449–458.

Rørth, P., 1996 A modular misexpression screen in *Drosophila* detecting tissue-specific phenotypes. Proc. Natl. Acad. Sci. USA **93:** 12418–12422.

Rørth, P., K. Szabo, A. Bailey, T. Laverty, J. Rehm *et al.*, 1998 Systematic gain-of-function genetics in *Drosophila*. Development **125:** 1049–1057.

Roth, S., D. Stein and C. Nusslein-Volhard, 1989 A gradient of nuclear localization of the dorsal protein determines dorsoventral pattern in the *Drosophila* embryo. Cell **59:** 1189–1202.

Roth, S., Y. Hiromi, D. Godt and C. Nusslein-Volhard, 1991 cactus, a maternal gene required for proper formation of the dorsoventral morphogen gradient in *Drosophila* embryos. Development **112:** 371–388.

Roth, S., F. S. Neuman-Silberberg, G. Barcelo and T. Schupbach, 1995 cornichon and the EGF receptor signaling process are necessary for both anterior-posterior and dorsal-ventral pattern formation in *Drosophila*. Cell **81:** 967–978.

Rubin, G. M., 1998 The *Drosophila* genome project: a progress report. Trends Genet. **14:** 340–341.

Rudkin, G. T., 1972 Replication in polytene chromosomes, pp. 59–85 in *Developmental Studies on Giant Chromosomes*, edited by W. Beermann. Springer-Verlag, Berlin.

Rusch, J., and M. Levine, 1997 Regulation of a *dpp* target gene in the *Drosophila* embryo. Development **124:** 303–311.

Russell, S. R. H., and K. Kaiser, 1993 mst35b, a male germline specific gene. Abstracts 13th Eur. Dros. Res. Conf.: I2.

Saccharomyces Genome Database, 1999 http://genome-www.stanford.edu/Saccharomyces/.

Sapir, A., R. Schweitzer and B. Z. Shilo, 1998 Sequential activation of the EGF receptor pathway during *Drosophila* oogenesis establishes the dorsoventral axis. Development **125:** 191–200.

Satoh, A. K., F. Tokunaga and K. Ozaki, 1997 Rab proteins of *Drosophila melanogaster*: novel members of the Rab-protein family. FEBS Lett. **404:** 65–69.

Schaeffer, S. W., and C. F. Aquadro, 1987 Nucleotide sequence of the *Adh* gene region of *Drosophila pseudoobscura*: evolutionary change and evidence for an ancient gene duplication. Genetics **117:** 61–73.

Schimmoler, F., E. Diaz, B. Muhlbauer and S. P. Pfeffer, 1998 Characterization of a 76kDa endosomal, multispanning membrane protein that is highly conserved throughout evolution. Gene **216:** 311–318.

Schmiedeknecht, G., C. Kerkhoff, E. Orso, J. Stoehr, C. Aslanidis *et al.*, 1996 Isolation and characterization of a 14.5-kDa trichloroacetic-acid-soluble translational inhibitor protein from human monocytes that is upregulated upon cellular differentiation. Eur. J. Biochem. **242:** 339–351.

Schupbach, T., and E. Wieschaus, 1986 Germline autonomy of maternal-effect mutations altering the embryonic body pattern of *Drosophila*. Dev. Biol. **113:** 443–448.

Schupbach, T., and E. Wieschaus, 1989 Female sterile mutations on the second chromosome of *Drosophila melanogaster*. I. Maternal effect mutations. Genetics **121:** 101–117.

Schweisguth, F., and J. W. Posakony, 1992 Suppressor of Hairless, the *Drosophila* homolog of the mouse recombination signal-binding protein gene, controls sensory organ cell fates. Cell **69:** 1199–1212.

Self, T., M. Mahony, J. Fleming, J. Walsh, S. D. M. Brown *et al.*, 1998 *Shaker-1* mutations reveal roles for myosin VIIA in both development and function of cochlea hair cells. Development **125:** 557–566.

Shen, W., and G. Mardon, 1997 Ectopic eye development in *Drosophila* induced by directed dachshund expression. Development **124:** 45–52.

Sigrist, S., G. Ried and C. F. Lehner, 1995 Dmcdc2 kinase is required for both meiotic divisions during *Drosophila* spermatogenesis and is activated by the twine cdc25 phosphatase. Mech. Dev. **53:** 247–260.

Simon, M. A., D. D. L. Bowtell, G. S. Dodson, T. R. Laverty and G. M. Rubin, 1991 Ras1 and a putative guanine nucleotide exchange factor perform crucial steps in signaling by the sevenless protein tyrosine kinase. Cell **67:** 701–716.

Smithies, O., G. E. Connell and G. H. Dixon, 1962 Chromosomal rearrangements and the evolution of haptoglobin genes. Nature **196:** 232–236.

Smoller, D. A., D. Petrov and D. L. Hartl, 1991 Characterization of bacteriophage P1 library containing inserts of *Drosophila* DNA of 75–100 kilobase pairs. Chromosoma **100:** 487–494.

Soehnge, H., X. Huang, M. Becker, P. Whitkey, D. Conover *et al.*, 1996 A neurotransmitter transporter encoded by the *Drosophila* inebriated gene. Proc. Natl. Acad. Sci. USA **93:** 13262–13267.

Sonnhamer, E. L., S. R. Eddy and R. Durbin, 1997 Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins **28:** 405–420.

Sorsa, V., 1988 *Chromosome Maps of Drosophila*, Vols. 1 and 2. CRC Press, Boca Raton, FL.

Sotillos, S., F. Roch and S. Campuzano, 1997 The metalloprotease-disintegrin Kuzbanian participates in Notch activation during growth and patterning of *Drosophila* imaginal discs. Development **124:** 4769–4779.

Spain, B. H., K. S. Bowdish, A. Pacal, S. Fluckiger Staub, D. Koo *et al.*, 1996 Two human cDNAs, including a homolog of Arabidopsis FUS6 (COP11), suppress G-protein- and mitogen-activated protein kinase-mediated signal transduction in yeast and mammalian cells. Mol. Cell. Biol. **16:** 6698–6706.

Spearman, C., 1904 The proof and measurement of association between two things. Am. J. Psychol. **15:** 72–101.

Spradling, A. C., and G. M. Rubin, 1981 *Drosophila* genome organization: conserved and dynamic aspects. Annu. Rev. Genet. **15:** 219–264.

Spradling, A. C., D. M. Stern, I. Kiss, J. Roote, T. Laverty *et al.*, 1995 Gene disruptions using P transposable elements: an integral component of the *Drosophila* genome project. Proc. Natl. Acad. Sci. USA **92:** 10824–10830.

Spradling, A. C., D. Stern, A. Beaton, E. J. Rhem, N. Mozden *et al.*, 1999 The BDGP gene disruption project: single P-element insertions mutating 30% of Drosophila autosomal genes. Genetics **153:** 135–177.

Stathakis, D. G., E. S. Pentz, M. E. Freeman, J. Kullman, G. R. Hankins *et al.*, 1995 The genetic and molecular organization of the Dopa decarboxylase gene cluster of *Drosophila melanogaster*. Genetics **141:** 629–655.

Sternberg, N., 1990 Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. Proc. Natl. Acad. Sci. USA **87:** 103–107.

Strobel, E., P. Dunsmuir and G. M. Rubin, 1979 Polymorphisms in the chromosomal locations of elements of the *412*, *copia* and *297* dispersed repeated gene families in *Drosophila*. Cell **17:** 429–439.

Stroumbakis, N. D., Z. Li and P. P. Tolias, 1996 A homolog of human transcription factor NF-X1 encoded by the *Drosophila* shuttle craft gene is required in the embryonic central nervous system. Mol. Cell. Biol. **16:** 192–201.

Sturtevant, A. H., 1925 The effects of unequal crossing over at the Bar locus in *Drosophila*. Genetics **10:** 117–147.

Styhler, S., A. Nakamura, A. Swan, B. Suter and P. Lasko, 1998 *vasa* is required for GURKEN accumulation in the oocyte, and is involved in oocyte differentiation and germline cyst development. Development **125:** 1569–1578.

Sun, X., J. Wahlstrom and G. Karpen, 1997 Molecular structure of a functional *Drosophila* centromere. Cell **91:** 1007–1019.

Tatei, K., H. Cai, Y. T. Ip and M. Levine, 1995 Race: a *Drosophila* homologue of the angiotensin converting enzyme. Mech. Dev. **51:** 157–168.

Taylor, C. A. M., D. Coates and A. D. Shirras, 1996 The *Acer* gene of *Drosophila* codes for an angiotensin-converting enzyme homologue. Gene **181:** 191–197.

Tessier-Lavigne, M., and C. S. Goodman, 1996 The molecular biology of axon guidance. Science **274:** 1123–1133.

Tolias, P. P., and N. D. Stroumbakis, 1998 The *Drosophila* zygotic lethal gene shuttle craft is required maternally for proper embryonic development. Dev. Genes Evol. **208:** 274–282.

Van Vactor, D., H. Sink, D. M. Fambrough, R. Tsoo and C. S. Goodman, 1993 Genes that control neuromuscular specificity in *Drosophila*. Cell **73:** 1137–1153.

Varshavsky, A., 1997 The ubiquitin system. Trends Biochem. Sci. **22:** 383–387.

Waldmann, R., and M. Lazdunski, 1998 H^+-gated cation channels: neuronal acid sensors in the NaC/DEG family of ion channels. Curr. Biol. **8:** 418–424.

Walter, M. F., L. L. Zeineh, B. C. Black, W. E. McIvor, T. R. Wright *et al.*, 1996 Catecholamine metabolism and in vitro induction of premature cuticle melanization in wild type and pigmentation mutants of *Drosophila melanogaster*. Arch. Insect Biochem. Physiol. **31:** 219–233.

Wang, Y., C. L. Farr and L. S. Kaguni, 1997 Accessory subunit of mitochondrial DNA polymerase from *Drosophila* embryos. Cloning, molecular analysis, and association in the native enzyme. J. Biol. Chem. **272:** 13640–13646.

Weil, D., S. Blanchard, J. Kaplan, P. Guilford, F. Gibson *et al.*, 1995 Defective myosin VIIA gene responsible for Usher syndrome type 1B. Nature **374:** 60–61.

Weinstein, J., F. W. Jacobsen, J. Hsu-Chen, T. Wu and L. G. Baum, 1994 A novel mammalian protein, p55CDC, present in dividing cells is associated with protein kinase activity and has homology to the *Saccharomyces cerevisiae* cell division cycle proteins Cdc20 and Cdc4. Mol. Cell. Biol. **14:** 3350–3363.

Welch, M. D., A. H. de Pace, S. Verma, A. Iwamatsu and T. Mitchison, 1997 The human ARP2/3 complex is composed of evolutionarily conserved subunits and is localized to cellular regions of dynamic actin filament assembly. J. Cell Biol. **138:** 375–384.

Whiteley, M., P. D. Noguchi, S. M. Sensabaugh, W. F. Odenwald and J. A. Kassis, 1992 The *Drosophila* gene escargot encodes a zinc finger motif found in snail-related genes. Mech. Dev. **36:** 117–127.

Wolfe, K. H., and D. C. Shields, 1997 Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387:** 708–713.

Woodruff, R. C., and M. Ashburner, 1979a The genetics of a small autosomal region of *Drosophila melanogaster* containing the structural gene for alcohol dehydrogenase. I. Characterization of deficiencies and mapping of *Adh* and visible mutations. Genetics **92:** 117–132.

Woodruff, R. C., and M. Ashburner, 1979b The genetics of a small autosomal region of *Drosophila melanogaster* containing the structural gene for alcohol dehydrogenase. II. Lethal mutations in the region. Genetics **92:** 133–149.

Wormpep, 1999 http: // www.sanger.ac.uk / Projects / C_elegans/ wormpep/.

Wright, T. R. F., 1987 The genetics of biogenic amine metabolism, sclerotization, and melanization in *Drosophila melanogaster*. Adv. Genet. **24:** 127–222.

Xie, Z., and D. H. Price, 1996 Purification of an RNA polymerase II transcript release factor from *Drosophila*. J. Biol. Chem. **271:** 11043–11046.

Xu, Y., G. Helt, J. R. Einstein, G. M. Rubin and E. C. Uberbacher, 1995 *Drosophila* GRAIL: an intelligent system for gene recognition in *Drosophila* DNA sequences, pp. 128–135 in *Symposium on Intelligence in Neural and Biological Systems*. IEEE Computer Society, Los Alamitos, CA.

Yagi, Y., and S. Hayashi, 1997 Role of the *Drosophila* EGF receptor in determination of the dorsoventral domains of escargot expression during primary neurogenesis. Genes Cells **2:** 41–53.

Yeung, K. C., J. A. Inostroza, F. H. Mermelstein, C. Kannabiran and D. Reinberg, 1994 Structure-function analysis of the TBP-binding protein Dr1 reveals a mechanism for repression of class II gene transcription. Genes Dev. **8:** 2097–2109.

Yeast Proteome Database, 1998 *The Yeast Proteome Handbook*. Ed. 5. Proteome Inc., Beverly, MA.

Zheng, W., G. Feng, D. Ren, D. F. Eberl, F. Hannan *et al.*, 1995 Cloning and characterization of a calcium channel α_1 subunit from *Drosophila melanogaster* with similarity to the rat brain type D isoform. J. Neurosci. **15:** 1132–1143.

Ziv, J., and A. Lempel, 1977 A universal algorithm for sequential data compression. IEEE Trans. Inf. Theory **23:** 337–343.

Communicating editor: T. C. Kaufman

APPENDIX: DETAILED DESCRIPTION OF GENES IDENTIFIED IN THE *Adh* REGION

B4: *B4* was discovered by Sotillos *et al.* (1997) and is the gene 823 bp distal to, and divergently transcribed from, *kuz*. The *P*-element insertion *PZ05337* is within *B4*. This mutation is viable and fertile with *Df(2L)b84a7*, an including deletion. The *P* element *k01405* [a cluster mate of *k01403*, Table S1 (http://www.genetics.org/cgi/content/full/153/1/179/DC1)] is a lethal *kuz* allele but may also affect *B4* function, since the viability of hemizygous *k01405* flies can be increased by *C765:GAL4* driving *UAS:B4* (Sotillos *et al.* 1997). *B4* corresponds to *BG:DS07660.4* (only the N terminus is on this sequence) and the predicted protein has no similarity to other proteins, even when the full-length protein of Sotillos *et al.* (1997) is used in a BLASTP analysis.

kuz (l(2)34Da): l(2)34Da was first identified as being a lethal associated with *TE34Ca*, an insertion of G. Ising's w^+rst^+ element, and its alleles *TE34Cb* and *TE34Cc* (M. Ashburner and J. Roote, unpublished observations). It is *kuzbanian*, encoding a disintegrin-like metalloprotease of the ADAM family (*BG:DS07660.3*; Fambrough *et al.* 1996; Rooke *et al.* 1996). *kuz* is required for *Notch* signal transduction, perhaps for the proteolytic cleavage of the *Notch* protein (Pan and Rubin 1997; Sotillos *et al.* 1997). Several *P*-element alleles of *kuz* are known, only some of which are lethal [see Table S1 (http://www.genetics.org/cgi/content/full/153/1/179/DC1)].

BG:DS07660.1: This gene is predicted to encode a protein of 453 amino acids that shows significant similarity in sequence to sodium/phosphate cotransporters of mammals (*e.g.*, BLASTP, $P = 10^{-59}$, 32% identity, over 88% of length, to the brain-specific sodium-dependent inorganic phosphate transporter of rat, SP:Q28722). It is also similar (30% identity over 86% of length) to a Na^+-dependent inorganic phosphate cotransporter of *D. melanogaster* mapped to 43BC (EMBL:Y07720). PSORT predicts that the protein has eight transmembrane domains, as do other members of this protein family (Griffith and Sansom 1998).

BG:BACR48E02.4: By virtue of significant sequence similarity with the human and mouse RAS-suppressor protein RSU1 (*e.g.*, BLASTP, $P = 10^{-73}$, 55% identity over 86% of its length with SP:Q15404), this predicted gene probably codes for a small GTPase regulatory/interacting protein similar to that identified in mice by Cutler *et al.* (1992) in an expression cloning assay for suppressors of the v-Ras phenotype.

BG:DS01368.1: The predicted protein product of this gene is weakly similar (BLASTP, $P = 10^{-20}$, 26% identity over 51%) to a hypothetical protein of *C. elegans* (*C26B9.1*, SPTREMBL:Q18202).

BG:DS08249.2: This gene almost certainly encodes a Drosophila mitochondrial glycerol 3-phosphate dehydrogenase, but it is not that known as *Gpo* (which maps to chromosome arm 2R; Davis and MacIntyre 1988). The protein product predicted for *BG:DS08249.2* has significant matches to the mitochondrial glycerol 3-phosphate dehydrogenase of organisms as different as human ($P = 10^{-105}$ with SP:P43304) and *Saccharomyces cerevisiae* ($P = 10^{-83}$ with SP:P32191) as well as having both PROSITE and PFAM flavin adenine dinucleotide (FAD)-dependent glycerol-3-phosphate dehydrogenase matches.

BG:DS08249.3: The product of *BG:DS08249.3* has a PROSITE (PS00518) and PFAM RING-finger domain (PF00097, $P = 7.9 \times 10^{-9}$) but the only significant BLASTP match is with a hypothetical human protein ($P = 10^{-75}$, 43% identity over 91% of its length with SPTREMBL:O75598). Weaker matches are seen with other C3HC4-type zinc finger proteins, *e.g.*, the Lnx protein of mouse ($P = 10^{-11}$, with SPTREMBL:O70623) and a hypothetical protein of *C. elegans* ($P = 10^{-9}$, with *F45G2.6*, SPTREMBL:O62248).

BG:DS00797.1: The *P* element *k07245* is a viable and phenotypically invisible insertion (although associated with a lethal chromosome) that is located 9 bp 5′ to the putative start of transcription of this gene. One out of 135 transposase-induced excisions of this *P* element is a long distally extending deletion (at least to *kuz*); this deletion is not mutant for *l(2)34Db*, giving a distal limit for this gene. The protein encoded by *BG:DS00797.1* is predicted to be a transmembrane domain protein (PSORT), similar to the EMP70 protein of *S. cerevisiae* (BLASTP, $P = 10^{-94}$, 34% identity over 72% of length) and a related protein from *Arabidopsis thaliana* (SPTREMBL:O04091). It has been suggested by Schimmoller *et al.* (1998) that members of the EMP70 protein family may be involved in small molecule transport in the endosome.

BG:DS00797.2: This hypothetical protein is similar to proteins from *Escherichia coli*, *S. cerevisiae*, and *Pennisetum ciliare*, whose functions are unknown but that belong to the same protein family (UPF0010).

p38b: This gene, encoding a MAP kinase (MAPK), corresponds to *BG:DS00797.3* as shown by its sequence. It was first found on our sequence by Han *et al.* (1998) and was also identified by an EDGP STS sequence (*ESTS:186F5S*). It is implicated in the antimicrobial response pathway, overexpression downregulating the induction of defense proteins by bacteria. *p38b* has also been identified by Adachi-Yamada *et al.* (1999) who have shown that it is involved in the TGFβ signalling pathway, because expression of a dominant-negative form causes a *dpp*-like phenotype and enhances the *dpp* mutant phenotype. *p38b* is very similar in protein sequence to human mitogen-activated protein kinase p38 (72% amino acid sequence identity). In *D. melanogaster*, there is a second p38 homolog, *Mpk2*, mapping at 95E. The proteins encoded by *p38b* and *Mpk2* (= *p38a*) are nearly 80% identical in amino acid sequence. Whether or not they can functionally substitute for each other is not yet known. *p38b* is the fourth MAPK to be identified in Drosophila; the others are the products of the *rolled* and *basket* genes, belonging to the ERK2 and JNK families of MAP kinases, respectively; both *p38b* and *Mpk2* belong to the stress-activated family.

BG:DS00797.4: The conceptual protein of this predicted gene only shows significant similarity with one of unknown function from *C. elegans*, *F26C11.1* (BLASTP, $P = 10^{-38}$ with SPTREMBL:Q17843, a protein with PROSITE histidine acid phosphatase signatures), and another of unknown function from the plant *Pimpinella brachycarpa* (BLASTP, $P = 10^{-37}$ with SPTREMBL: O81652).

BG:DS00797.5: The predicted protein of *BG: DS00797.5* has a PFAM ABC transporter pattern ($P = 1.9 \times 10^{-40}$) and shows BLASTP similarities in its C-terminal exon with ABC transporters from mammals, but the identities are relatively low (∼32%). At a similar level of identity, it resembles a hypothetical protein of *C. elegans*, *F33E11.4*, which also belongs to the ABC transporter protein family.

BG:DS00797.6: The protein of this predicted gene shows significant similarities with only two others: one is a hypothetical protein of *C. elegans*, *K09A11.1*, said to be similar to transposases ($P = 10^{-14}$, 21% identity over 39% of residues with SPTREMBL:Q21374) and the other is the transposase of the Hermit element of *Lucilia cuprina* ($P = 10^{-11}$, 19% identity over 66% of residues with SPTREMBL:Q25239).

anon-34Da: This gene was named for transcript 7 of Bonfini *et al.* (1992), mapping ∼20 kb distal to *Sos*. From its position, it probably corresponds to *BG:DS00797.7*, and this may correspond to *l(2)34Db*. The predicted protein is similar to the SEC7 protein of *S. cerevisiae* ($P = 10^{-171}$, 33% identity over 28% of length). In yeast this protein is essential for vegetative growth and is involved in endoplasmic reticulum to Golgi protein transport (Achstetter *et al.* 1988; Franzusoff *et al.* 1991). The Drosophila protein also shares a domain with the bovine guanyl-nucleotide exchange protein (Moringa *et al.* 1996; 61% identity over 46% of length) and a similar protein is found in *A. thaliana* (*F23E12.60*).

BG:DS00941.1: This is a Drosophila carbonate dehydratase. It shows highly significant BLASTP matches over its entire length with this enzyme from human, mouse, Chlamydomonas, Anabaena, and zebra fish, and there is a similar sequence predicted in *C. elegans* (*R173.1*). In vertebrates there are several carbonate dehydratases with different subcellular localizations. *BG:DS00941.1* is most similar to the human *CA7* and mouse *Car2* genes, known or presumed to code for cytosolic forms of the enzyme, which catalyzes the hydration of carbon dioxide. There is biochemical evidence for three carbonate dehydratase genes in Drosophila (Choudhary *et al.* 1992), but these genes had not been characterized at the molecular level.

BG:DS00941.2: This gene would appear to code for one of two Drosophila RNA adenosine deaminases (Bass 1997). The other is also a predicted gene from the EDGP (*EG:BACN35H14.1*). The protein predicted from *BG:DS00941.2* shows ∼30% identity over its entire length to the double-stranded RNA adenosine deaminases of human, mouse, and Xenopus, which are involved in pre-mRNA editing. There are equally similar proteins predicted for *S. cerevisiae* (*YGL243W*), *S. pombe*, and *C. elegans* (*T20H4.4*) that lack double-stranded RNA-binding domains, and the *S. cerevisiae* protein has been shown to be an adenosine deaminase acting on tRNA (ADAT).

BG:DS00941.2 was independently identified as an RNA adenosine deaminase by L. Keegan (personal communication), who has named it *Adat*. The absence of a dsRNA-binding domain from this protein, and *in vitro* studies of the expressed protein, have led L. Keegan and colleagues (personal communication) to the conclusion that this protein functions as a tRNA, rather than as a pre-mRNA, adenosine deaminase. This gene probably does not correspond to *l(2)34Db*, because we expect *BG:DS00941.2* to have been included in a 15-kb *Kpn*1 *Sos* transgene (Bonfini *et al.* 1992) that does not rescue alleles of this lethal locus.

BG:DS00941.3: The only significant BLASTP match with the protein predicted for *BG:DS00941.3* is to a human cDNA sequence (SP:O43351) that matches the human EST EMBL:AA085966, itself said to be similar to the human P31 proteasome subunit ($P = 10^{-10}$, 53% identity over 21% of length).

Sos (l(2)34Ea): l(2)34Ea was one of the most mutable genes in the early EMS mutagenesis experiments. It is the gene named *Son of sevenless* by Rogge *et al.* (1991), who recovered an allele as a dominant suppressor of a gain-of-function allele of *sevenless*. The same gene was identified as an enhancer of *sevenless* by Simon *et al.* (1991), who showed it to encode a guanine-nucleotide exchange factor required for signal transduction in the RAS pathway (see also Bonfini *et al.* 1992). *Sos* corresponds to *BG:DS00941.4*, as is shown by direct sequence comparison.

black: The first mutant allele of the *black* body color gene was discovered by T. H. Morgan in October 1910. It is a nonvital gene and all mutant alleles result in very darkly pigmented adult flies and white pupal cases. The phenotype results from a failure to synthesize β-alanine (Hodgetts 1972) and can be corrected by dietary β-alanine (Jacobs 1974). β-alanine forms an adduct with dopamine (Wright 1987) and this is required for proper tanning of the cuticle (the β-alanyl-dopamine synthetase is probably the product of the *ebony* gene; see Walter *et al.* 1996). There are two possible pathways of β-alanine synthesis, by decarboxylation of aspartic acid and by pyrimidine catabolism (Jacobs 1974). The facts that *black* mutant alleles are enhanced by mutations in *su(r)*, which encodes the NAD^{+}-dependent dihydrouracil dehydrogenase (Bahn 1972), and that 6-azathymidine produces a *black* phenocopy (Pedersen 1982) suggested that pyrimidine catabolism is the more important in Drosophila.

The predicted gene *BG:DS00941.5* maps between *Sos* and *BG:DS00941.6*; we argue that the latter is *l(2)34Dc* (see below). This is precisely the genetic location of *black* by deletion mapping; moreover, these three genes are so very closely spaced that we can be confident that no others are to be found in this 18-kb interval. *BG: DS00941.5* shows a good match (45–50% identity) to glutamate decarboxylase from mammals (mice, human) and to the rat cysteine sulfinate decarboxylase (SPTR-EMBL:Q64611). The *Drosophila* gene had been sequenced by Phillips *et al.* (1993). A cDNA of this gene, the gift of M. Phillips, crosses the breakpoint of *Tp(2;3)b79d6*, an aberration allele of *black* that is viable when hemizygous with long deletions of the *black* region. There is a second gene encoding glutamate decarboxylase in Drosophila, which is required for the synthesis of the neurotransmitter γ-aminobutyric acid, *Gad2*, mapping to 64A (Jackson *et al.* 1990). Glutamate decarboxylase is known to have aspartate decarboxylase (ADC) activity in mammals (Porter and Martin 1988). This suggests that the absence of β-alanine in *black* mutations is due to a failure of aspartic acid catabolism, rather than of pyrimidine breakdown, despite the data of Jacobs (1974) that indicated no difference in the decarboxylation of ^{14}C-aspartic acid between a *black* strain and a wild type (see also Phillips *et al.* 1993).

tamas (l(2)34Dc): This was identified as a lethal locus from eight EMS-induced alleles. Adult escapers have missing bristles on the head and notum and blistered wings with some disruption of the wing veins. This gene has been deletion mapped to between *black* and *l(2)34Dd* or *l(2)34Df* (the last two genes have not been ordered genetically). Because *l(2)34Dd* is a Drosophila homolog of yeast *SOP2* (below; *BG:DS00941.7*) and because *BG:DS00941.6* is the only open reading frame between *black* and *Sop2* in a very closely packed interval, we conclude that *l(2)34Dc* is *BG:DS00941.6, i.e.*, encodes the catalytic subunit of the mitochondrial DNA polymerase, previously sequenced from Drosophila by two groups (Lewis *et al.* 1996; Ropp and Copeland 1996). It is interesting that *BG:DS00941.9*, just 8 kb proximal, encodes the accessory subunit of this enzyme (see below).

B. Iyengar, J. Roote and A. R. Campos (unpublished results) identified an EMS-induced mutation of *l(2)34Dc* in a screen for larvae defective in their response to light. This phenotype was found to be a consequence of a defect in larval locomotor behavior. Four mutant alleles of *l(2)34Dc*, which they call *tamas*, were sequenced; two were missense mutations and the others small (1-bp and 5-bp) deletions within the coding region of the gene encoding the catalytic subunit of the mitochondrial DNA polymerase.

Sop2 (l(2)34Dd): This gene is known only from three EMS-induced lethal alleles. Hudson and Cooley (1998) have shown, by transformation rescue, that these are in *BG:DS00941.7*, a Drosophila homologue of the *Schizosaccharomyces pombe* Suppressor of Profilin 2 (*SOP2*) gene. A similar sequence is the 41-kD subunit of the human ARP2/3 complex, a protein complex involved in the control of actin-filament assembly (Welch *et al.* 1997).

Orc5 (l(2)34Df): Only two EMS-induced lethal alleles are known for *l(2)34Df*. Genetically, *l(2)34Df* maps between *l(2)34Dc* and *l(2)34Dd* or between *l(2)34Dd* and *l(2)34De*, and there are two candidate-predicted genes: *BG:DS00941.8* and *BG:DS00941.9*. The former, *BG:DS00941.8*, encodes the Drosophila Origin Recognition Complex subunit 5 protein (Gossen *et al.* 1995) and, as shown by M. Pflumm (personal communication) by transformation rescue, is *l(2)34Df*. This conclusion places *l(2)34Df* between *l(2)34Dd* and *l(2)34De*.

MtpolB (l(2)34De): Genetically, *l(2)34De* maps between *l(2)34Dd* (*Sop2*) or *l(2)34Df* (*Orc5*) and *l(2)34Dg* (*RpII33*). The evidence for the gene order *l(2)34De l(2)34Dg* comes from complementation data with *T(2;3)b89e12*, which is *l(2)34Dd*$^-$ *l(2)34Df*$^-$ *l(2)34De*$^-$ *l(2)34Dg*$^+$. There is only one predicted gene in the 1.9 kb separating *Orc5* and *RpII33*, which is the gene encoding the accessory subunit of the mitochondrial DNA polymerase (*BG:DS00941.9*; Wang *et al.* 1997). It is a reasonable hypothesis that *l(2)34De* encodes this protein.

RpII33 (l(2)34Dg): l(2)34Dg was first identified from two EMS-induced lethal alleles; subsequently the *P*-element insertion *k05605* was shown to be allelic. This insertion is in the 5′ of *BG:DS00941.10*, encoding a homolog to the 33-kD subunit of RNA polymerase II from mammals, *S. cerevisiae*, and *A. thaliana*; we can be confident that this is indeed the *RpII33* gene of Drosophila, because the amino acid identities are ∼68% between the entire Drosophila protein and its human homolog.

BG:DS08220.1: This is a predicted gene with a match to human and *C. elegans* EST sequences of unknown function. The *P* elements *PZ06646* and *rN149* are phenotypically silent insertions at the same nucleotide 1 kb upstream of this transcription unit; the viable insertion *k10802* is inserted 11 bp 5′ to this transcription unit. Over 180 transposase-induced excisions of the *PZ06646* element have been recovered; all are viable when heterozygous with long deletions of the 34D-35B interval. Three (of 84) transposase-induced excisions of *rN149* are associated with lethal mutations, two of which map distal to *BG:DS08220.1* and, presumably, are due to secondary events, and the third of which deletes *Ance-wb*. The product of *BG:DS08220.1* may well be involved in a signal transduction pathway. The most similar proteins are the hypothetical *KIAA0167* human protein (BLASTP, $P = 10^{-148}$, 42% identity over 51% of residues) and hypothetical *C. elegans* protein *Y39A1A.15B* (BLASTP, $P = 10^{-139}$, 45% identity over 30% of residues), but significant similarities are seen over short regions with the pig and rat inositol 1,2,3,4-tetrakisphosphate receptor (or binding protein).

anon-34Ea: This gene was defined by FlyBase for a transcript immediately 5′ to *Ance* detected by Tatei *et al.* (1995). It is *BG:DS08220.2*, and is without any significant database matches. The 16.5-kb *Eco*RI fragment transformed by Tatei *et al.* (1995) carries *anon-34Ea* (and *Ance*) and rescues mutant alleles of *Ance*, as well as the homozygously deleted region in *Df(2L) b88f32/Df(2L)nBR55* heterozygotes (Tatei *et al.* 1995). The viable insertion *EP(2)2171* is inserted within the first exon of this gene.

Ance (l(2)34Eb): This vital gene was identified by two EMS-induced alleles. It was shown by transformation rescue to encode a peptidyl-dipeptidase A, similar to human angiotensin-converting enzyme, hence *Ance*, by Tatei *et al.* (1995). It is *BG:DS08220.3*, and was also sequenced by Cornell *et al.* (1995), but mismapped by them to 34A. *Ance* protein is an early marker for amnioserosal differentiation (Tatei *et al.* 1995; Frank and Rushlow 1996), where it is activated by the *zen* homeodomain transcription factor (Rusch and Levine 1997). There is a second gene encoding an angiotensin-converting enzyme-like protein in Drosophila, *Acer*, mapping at 29D (Taylor *et al.* 1996). Clearly, these are not functionally redundant; indeed, Houard *et al.* (1998) show that the purified ANCE and ACER enzymes, which are 47% identical in amino acid sequence, have different substrate specificities and expression patterns.

Acyp: A. Bairoch identified a sequence encoding a homolog of vertebrate acylphosphatase in our sequence of DS00180; this is *BG:DS00180.1* (SP:P56544). Biochemical studies of the protein expressed in *E. coli* confirm its function (Pieri *et al.* 1998).

BG:DS00180.2, BG:DS00180.3: BG:DS00180.2 and BG:DS00180.3 are predicted genes whose protein sequences are 28% identical and have valine/proline-rich repeats. These proteins have significant database matches in unfiltered BLASTP to articulins, cytoskeletal proteins of the epiplasm of flagellates and ciliates. Articulins are characterized by VPVPxxVxxxV repeats (Marrs and Bouck 1992). *BG:DS00180.2*, *e.g.*, has four copies of a VIK[K|E]V[P|H]VPV motif and four copies of a PVEKx-[V|I]HVPV[H|K]V motif.

BG:DS00180.5: The protein of this predicted gene has a limited region of similarity with angiotensin-converting enzymes from mammals and Drosophila, *e.g.*, 42% identity over 13% to the human DCP1 protein (SP:P12821). It does not have a PROSITE zinc metallopeptidase, zinc-binding region signature, nor is it similar overall with either the *Ance* or *Acer* proteins. The existence of this gene is based on *ab initio* prediction; it has no EST matches.

BG:DS00180.12, BG:DS00180.7, BG:DS00180.8, BG:DS-00180.9, BG:DS00180.10, and BG:DS00180.14: These are a cluster of predicted genes, all of which show features of extracellular protein domains, such as EGF repeats and similarities to vertebrate tenascins and fibrillins. *Inter se* their similarities are in the twilight zone (18–28% identity) except for *BG:DS00180.12* and *BG:DS00180.8* (37% identity). Four of these genes have Drosophila EST sequences. Their relationships and structures require further study.

BG:DS00180.11: This is one of the two genes in this region that encode cytochrome P450s [the other is *l(2)35Fb*]. The most similar protein is *Cyp28a1* of *D. mettleri* (68% identity), one of a new family of cytochrome P450s identified as being induced by isoquinoline alkaloids found in the cactus hosts of this desert species (Danielson *et al.* 1997).

rk: rickets was discovered after UV mutagenesis by Edmondson (1948). All alleles cause a recessive visible phenotype characterized by bent legs (especially those of the metathorax) and unexpanded wings (at least in strong alleles). It is not lethal, because overlapping deletions [*e.g.*, *Df(2L)el80f1/Df(2L)b85f1A*] are viable (and extreme rickets). There is one *P*-element allele known, rk^{11P}; its insertion site maps some 4 kb upstream of the *rk* sequence as identified by J. Baker (personal communication) as corresponding to *BG:DS00180.13*. This gene encodes a 7TM protein that may be a neuropeptide hormone receptor because it shows sequence similarity to the mammalian G-protein-coupled lutropin-choriogonadotrophic hormone receptor (BLASTP, $P = 10^{-91}$ with SP:P22888; J. Baker, personal communication). The *rickets* protein is also similar in sequence to the product of the Drosophila *Fsh* gene, described as being related to the mammalian glycoprotein hormone receptors (Hauser *et al.* 1997).

BG:DS01514.2 and BG:DS05899.1: These genes are of rather different structure. The former has seven exons and the latter two. Yet their predicted proteins are of similar length (668 and 681 amino acids, respectively) and 43% identical (71% similar) in sequence. Both show significant similarities with long-chain-fatty-acid-CoA-ligases from species as different as *Archaeoglobus fulgidus*, yeasts, and mammals, and with similar genes in *C. elegans* (*R09E10.3*) and *A. thaliana* (*T08I13.8*). This is presumably their function in Drosophila. The *P* element *k09909* maps to *BG:DS01514.2*. M. Leptin and C. Coelho (personal communication) have sequenced cDNAs for both of these genes.

l(2)34Fa: This vital gene is known from two EMS alleles and one *P*-element insertion (*k00811*). The insertion site of the latter has been sequenced and falls 1.4 kb 5′ to the open reading frame of *BG:DS05899.2*. The predicted product of this gene has no sequence matches.

BG:DS05899.7: The predicted protein of this gene shows similarities to a variety of proteins from *C. elegans*, *S. cerevisiae*, Arabidopsis, and mammals. These all have leucine-rich repeats in common with *BG:DS05899.7*.

BG:DS05899.3: The product of *BG:DS05899.3* is cysteine rich and has relatively low similarities (BLASTP expectations in the range $P = 10^{-9}$ to 10^{-12}) with mammalian fibrillin 1 precursors as well as with the *apx-1* gene product of *C. elegans*. The latter is a *Delta*-like protein expressed maternally in the worm and interacting with the *glp-1* protein (a homolog of Drosophila *Notch*) in the determination of the anterior-posterior axis of the four-cell embryo (Mello *et al.* 1994).

BG:DS05899.4: This gene is predicted to encode a nicotinic acetylcholine receptor alpha chain. It shows 54% identity (over 57% of its length) with the human neuronal nicotinic acetylcholine receptor alpha-7 chain precursor (CHRNA7, SP:P36544) and its homologs in chicken and mouse. Three other nicotinic acetylcholine receptor alpha chains are known in Drosophila, two in 96A on chromosome arm *3R* and one at 7E on the *X* chromosome (data from FlyBase).

BG:DS01523.2: BG:DS01523.2 is predicted to encode a protein that has relatively low similarity (25% identity) to Drosophila *midline fasciclin* and fasciclin-like proteins from chick (SPTREMBL:O42390), mouse (osteoblast specific factor 2, SPTREMBL:Q62009), and a human TGFβ-induced protein (SP:Q15582). The C-terminal region of this 1894-residue predicted protein is very threonine rich (overall the predicted protein is 17.6% threonine), with many small repeat motifs, *e.g.*, nine copies of TT[P|R|N]APTTT[D|E|K], plus many small repeats (*e.g.*, five copies of TTTTA, four of TTTTS, four of EITTT).

smi35A: smi35A was identified by Anholt *et al.* (1996) on the basis of the reduced avoidance to benzaldehyde and other noxious chemicals associated with a *P*-element insertion. R. Anholt (personal communication) has discovered that a similar phenotype is associated with the insertions *k16716* and *k06901*. Both these and the original *smi35A* insertion map within a 21-bp interval some 12 kb 5′ to *wb*. Indeed, *k16716*, but not the other two insertions, is associated with a very weak wing-blister phenotype (when hemizygous with a wb^- deletion). However, the strongest smell-impaired phenotype is associated with the insertion *k11509*, which maps some 30 kb more distally, within the 5′ exon of *BG:DS01523.3* (R. Anholt, personal communication). One (of 128) transposase-induced loss of this element is a lethal allele of *wb*. This predicted gene encodes a YAK1/DYRK family protein kinase; the Drosophila and human proteins (DYRK2, SPTREMBL:Q92630) are 56% identical over 44% of the length of the former.

wb (l(2)34Fb): Alleles of *wing blister* are the most common lethals in EMS screens against deletions uncovering the *Adh* region. The alleles vary from being completely lethal to viable, with adult flies having a characteristic blister in the central wing. Several *P*-element alleles have been sequenced, some of which are lethal alleles and some viable. A lethal insertion, *PZ09437*, maps within a long intron of *BG:DS03792.1*, a gene encoding a protein similar to both laminin α-1 and α-2 chains of mouse and human. This gene has also been studied by Martin *et al.* (1999), who have determined both its molecular structure and its expression. The gene is among the largest in the *Adh* region, over 70 kb in length with a predicted mRNA of 10.8 kb spliced with at least 16 exons. Its size presumably accounts for its mutability, not only with EMS but also after irradiation; three chromosome aberrations are associated with *wb* alleles [*T(2;3)6r28*, *In(2LR)DTD121*, and *T(2;3)H68*]. There is an independent gene prediction included within *wb*, *BG:DS03792.2*.

BG:DS01068.10: This is one of several predicted genes to encode a serine protease. The protein of *BG:DS01068.10* is similar to trypsins from several organisms, from *Streptomyces glaucescens* to macaque. It is most similar to the theta-trypsin of *D. melanogaster* (37% identity over its entire length).

BG:DS01068.6: This is another gene encoding a protein conserved between yeasts and flies, but all of whose significant matches are themselves hypothetical. PSORT strongly predicts this protein to be nuclear. The matches are to *F32E10.1* of *C. elegans* (45% identity over 77% of residues), *YGR145W* of *S. cerevisiae* (38% identity over 76% of residues), and *SPCC330.09* of *S. pombe* (37% identity over 78% of residues). Mammalian EST matches indicate that a similar gene (or genes) will be found in mouse and human in due course.

Rab14: Rab14 is one of many genes in *D. melanogaster* encoding RAS-related proteins. By direct sequence comparison *BG:DS01068.7* is *Rab14*, which had been sequenced by Satoh *et al.* (1997) but mapped by them to 36A-B. This gene is also identified by an STS sequence derived from a cosmid mapped to 34F-35A (*ESTS:57H4T*, EMBL:Z50609). The phenotypically silent *P*-element insertion *k08712* is inserted at the 5′ end of *Rab14*. Three transposase-induced excisions of *k08712* are lethal (of 37 recovered). One is an allele of *l(2)35Aa* and two are alleles of *l(2)34Fd*. The lethality of the $l(2)35Aa^-$ derivative of *k08712* is rescued by a *P*-element insertion carrying a 5-kb *l(2)35Aa* rescue fragment (given to us by C. Flores) in the *Df(2L)k08712-rv21*/*Df(2L)TE35B-7* heterozygote, which is deleted for *Rab14*, *l(2)35Aa*, *spel1*, and *ppk*. These data suggest that *l(2)34Fd* is distal to *Rab14*, and that *Rab14* itself is not a vital gene.

l(2)35Aa: Seven EMS-induced lethal alleles of *l(2)35Aa* are known. *l(2)35Aa* corresponds to *BG:DS01068.8*, which encodes a protein similar to a polypeptide N-acetylgalactosaminyltransferase of human (SPTREMBL:Q10471), as was demonstrated by Flores and Engels (1999) by transformation rescue of mutant alleles and the overlapping deletions *Df(2L)b84hl* and *Df(2L)TE35B-7*.

spel1: spellchecker-1 encodes a Drosophila protein probably involved in DNA mismatch repair, because it carries a mutS protein family signature (Flores and Engels 1999). It corresponds to *BG:DS01068.9*. *spel1* is not a vital gene because a 5-kb *l(2)35Aa* transgene rescues the lethality of overlapping deletions [*Df(2L)TE35B-7*/*Df(2L)b84h1*] that are homozygously deleted for both *l(2)35Aa* and *spel1* (Flores and Engels 1999).

ppk: pickpocket encodes a protein whose sequence shows it to be a member of the DEG/ENaC protein superfamily (Adams *et al.* 1998; Waldmann and Lazdunski 1998). Adams *et al.* (1998) suggest that this may be involved as an ion channel protein in mechanosensory signal transduction, because it is expressed in a subset of multidendritic neurons. It corresponds to *BG:DS06238.1*. It is not a vital gene because the deletions *Df(2L)A400* and *Df(2L)b88h49* both remove *ppk* (M. Anderson, personal communication) and these deletions are viable when heterozygous with each other (see also above). This gene has also been sequenced by Darboux *et al.* (1998) and described as a multidendritic neuron sodium channel protein.

elbow (el) and pupal (pu): The genetics of the *elbow-no ocelli* region have long been known to be complex (see Davis *et al.* 1997). *elbow* and *pupal* have been known for many years, although, until the genetic analysis of the *Adh* region began, only a single allele of *elbow* had been recovered. The complex complementation patterns between the many alleles of *elbow* that have now been analyzed suggest that this "gene" is in fact two, *elB* and *elA*, and that mutations of each can act as dominant enhancers of mutations of the other. The insertion *EP(2)2039* is a weak *elbowB* allele, and enhances *Sco*, as do other alleles of *elB*; transposase-induced excisions of this element either revert the elbow phenotype, remain *elbow*, or are deletions extending proximal-ward (to include *pupal*) or distal-ward (to include *l(2)35Aa*). This *P* element is inserted at the 5′ end of the GENSCAN prediction for *BG:DS06238.3*, encoding a Zn-finger protein. We suggest that this gene is *elB*. If *BG:DS06238.3* is *elB*, then *BG:DS06238.4*, predicted to encode a protein with similarity to a Drosophila pupal cuticle protein (60% identity over 28% of length with the *Edg84A* protein), is probably *pupal* (whose most obvious phenotype is a failure of wing expansion), and *BG:DS08340.1* is probably *elA*. *BG:DS08340.1* is wholly contained within the 20-kb deletion associated with *el^1*; its sequence has no significant database matches. Although *elB*, *pu*, and *elA* are all nonvital individually, deleting all three genes results in pharate adult lethality, the adult escapers having crippled legs.

noc: no-ocelli was first identified by the absence of ocelli in certain viable overlapping deletion heterozygotes (Ashburner *et al.* 1982a). Subsequently, a number of viable alleles were found, including one associated with G. Ising's *w^+ rst^+* TE, *TE146* (now *TE35B*; Gubb *et al.* 1985). A lethal complementation group, described as *l(2)35Ba*, was clearly associated with *noc*, because heterozygotes between the EMS-induced lethal alleles of this group and viable *noc* alleles had no ocelli. In fact, this lethal locus and *noc* are the same gene, the viable alleles all being in 3′ regulatory regions (Chia *et al.* 1985; McGill 1985; Davis *et al.* 1990; Cheah *et al.* 1994). Three of the EMS-induced alleles die as embryos, showing a failure of embryonic head involution with hypertrophy of the supraesophageal ganglion (Cheah *et al.* 1994). Paradoxically, overlapping deletions for *noc* die as larvae, with no central nervous system phenotype; these three EMS alleles are recessive antimorphs (see discussion in Cheah *et al.* 1994). *noc* encodes a protein with a C2H2-like zinc finger and several long poly-alanine runs and corresponds to *BG:DS04641.1*, as shown by direct sequence comparison with the data of Cheah *et al.* (1994). This protein shows sequence similarity with the human SP1 and SP2 transcription factors.

noc shows complex genetic interactions with mutations at the *elA* and *elB* loci (Davis *et al.* 1997). It therefore is of some interest that *BG:DS06238.3*, which we suggest is *elB* and maps ~100 kb distal to *noc*, encodes a zinc finger protein showing 27% amino acid sequence identity with the *noc* protein.

BG:DS01486.1: Ubiquitin-protein ligases are required for the ubiquitination of proteins destined for breakdown via the 26S proteasome. *BG:DS01486.1* is the 12th gene in this family to be discovered in *D. melanogaster* (data from FlyBase); there are at least 13 in *S. cerevisiae* (Saccharomyces Genome Database 1999) and at least 10 in *C. elegans* (Wormpep 1999). *BG:DS01486.1* shows high identities (up to 83%) with 17-kD ubiquitin-conjugating enzyme E2 of organisms from yeast (UBC13p) to human (Varshavsky 1997).

osp. outspread was first recognized in Cambridge by the outspread wing phenotype of certain viable overlapping deletion heterozygotes (Woodruff and Ashburner 1979a). Subsequently, E. H. Grell (cited in Lindsley and Zimm 1992) identified an EMS-induced allele and many have been found since. It is not a vital gene, as complete deletions of *osp* are viable. Molecular mapping of aberration breakpoints associated with *osp* alleles on a phage chromosome walk showed that three mapped distal to *Adh* and four mapped proximal to this gene, leading to the conclusion that *Adh* was contained within *osp* (Chia *et al.* 1985). Subsequent work (McNabb *et al.* 1996) strengthened this hypothesis and, from our cDNA sequencing, we found that coding exons of *osp* map both distal to *Adh* and proximal (*BG:DS01486.7*). *Adh* and *Adhr* appear not to be the only genes included within *osp*; in addition to these are two transposable elements (*roo* and *jockey*) and two predicted genes, *BG:DS07721.1* and *BG:DS09219.1*. The second of these would be transcribed in the same direction as *osp*, and may be part of *osp* itself, if *osp* has an alternative transcript that has not yet been found as a cDNA (we already know of alternative transcripts of this gene that differ in their 3′ exons). *BG:DS07721.1* cannot be part of *osp* because it would be transcribed from the opposite strand (its existence is predicted by an EST sequence).

There are two *P*-element insertions in the 5′ exon of *osp*: one (*rJ571*) causes an *osp* phenotype, the other (*k13218*) does not. (A minority of transposase-induced excisions of *k13218*, 10 out of 225, are phenotypically outspread.) The gene is the largest we have found in the sequenced region, extending over 95 kb, with 5.3- and 3.9-kb cDNAs.

The predicted *osp* protein has a pleckstrin homology (PH) domain (PFAM:PF00169), implicating a role in the cytoskeleton. It shows some similarity to a protein involved in the control of the actin cytoskeleton in mice (p116Rip, SPTREMBL:P97434), to the myosin heavy chain products of the human *MYH3* and *MYH8* genes, and to the *S. cerevisiae* gene product *USO1* involved in intracellular protein transport.

Adh and Adhr: These are a pair of related genes, coding for proteins with 33% amino acid identity. The positions of the two introns that interrupt the coding regions of each are the same in the two genes, supporting the hypothesis that they arose by tandem duplication (Schaeffer and Aquadro 1987). The transcript of *Adhr* is much rarer than that of *Adh* and is always found as an *Adh-Adhr* dicistronic mRNA (Brogna and Ashburner 1997). These genes correspond to *BG:DS01486.8* and *BG:DS01486.9*, respectively. Despite its sequence matches *Adhr* is probably not an alcohol dehydrogenase; it is not an essential gene (Ashburner 1998).

BG:DS00810.1: The product of this predicted gene has a significant BLASTP score ($P = 10^{-19}$, 34% identity over 46% of length) to a hypothetical protein of *C. elegans* (*ZK652.6*).

BG:DS06874.2: High BLASTP scores ($P = 10^{-60}$, 39% identity over 94% of length) identify the product of *BG:DS06874.2* as being involved in a G-protein signal transduction pathway, because it is similar to the human protein GPS1 (and its rat homolog) isolated as a cDNA that suppresses gain-of-function mutations in the pheromone response pathway of *S. cerevisiae* and the RAS pathway in mammalian cells (Spain *et al.* 1996). The mammalian protein, and its Drosophila homolog, are also similar to the *FUS6* protein of *A. thaliana*, which is a negative regulator of light-mediated signal transduction (Castle and Meinke 1994).

BG:DS06874.3: The protein predicted to be the product of *BG:DS06874.3* has a PROSITE ATP/GTP-binding site motif A (P-loop) and PROSITE AAA-protein family signature. Its closest sequence match in the yeast genome is *MSP1*, encoding an AAA family ATPase of the inner mitochondrial membrane presumed to be involved in protein sorting (Nakai *et al.* 1993; $P = 10^{-63}$, 42% identity over 77% of its length). There are similar proteins in *C. elegans* (*K04D7.2*), *A. thaliana* (*T14P8.7*), and human (*SKD1*), and the *BG:DS06874.3* protein shows 36% amino acid sequence identity with the *TER94* gene product of *D. melanogaster*, isolated as a homolog of the yeast *CDC48* protein (Pinter *et al.* 1998). The *CDC48* protein is an essential AAA-family ATPase required for membrane fusion (Yeast Proteome Database 1998). The AAA family ATPases are a functionally diverse group of proteins, many of which are associated with the membranes of cell organelles (Patel and Latterich 1998). The predicted protein of *BG:DS06874.3* has a long C-terminal coiled-coil domain (PSORT prediction).

BG:DS06874.4 and BG:DS06874.6: The predicted protein products of these genes are 45% identical in amino acid sequence, and both products show significant similarities with a variety of serine proteases from organisms as different as *C. elegans* and human. These are not vital genes, because the heterozygote between the deletions *Df(2L)A72* and *Df(2L)A47* that removes both of these genes, is viable (J.-M. Reichhart, personal communication).

BG:DS03431.1: We predict that the protein product of *BG:DS03431.1* is a cation-dependent amino acid transporter. It shows 31% amino acid identity with the Drosophila *inebriated* protein (a Na^+/Cl^--dependent neurotransmitter transporter; Soehnge *et al.* 1996), and similar identities with Na^+/Cl^--dependent transporters from human (*SLC6A6*, a taurine transporter), *Manduca sexta* (*KAAT1*, amino acid transporter), rat (*SLC6A11*, GABA transporter), and even *Methanococcus jannaschii* (*MJ1319*, a putative sodium-dependent transporter). As expected for a protein of this function the *BG:DS03431.1* product is predicted by PSORT to have 12 transmembrane domains.

Mst35Ba and Mst35Bb: These are a tandem pair of related genes that encode protamine-like proteins (Russell and Kaiser 1993). They are probably not vital because *Df(2L)TE35D-5/Df(2L)TE35B-9* and *Df(2L)TE-35B-9/Df(2L)osp29* survive but are male sterile, suggesting that one or both of these may be required for male fertility. This conclusion is tentative, because these deletions remove much more than just these two *Mst* genes, but we reserve the symbol *ms(2)35Bi* for the genetic factor(s) responsible for this sterility. These protamine-like genes correspond to *BG:DS03431.2* and *BG: DS03431.3*, respectively.

BG:DS03144.1: This is a large predicted gene (~13.5 kb) with 11 predicted exons. Significant BLASTP matches are seen with a number of poorly characterized putative glycosyl phosphatidyl inositol (GPI)-anchored membrane-bound proteins with immunoglobulin-like domains [*e.g.*, the *D. melanogaster Amalgam* protein and locust lachesin ($P = 10^{-29}$ with SP:Q26474; Karlstrom *et al.* 1993)].

BG:DS03323.1: The *BG:DS03323.1* protein shares a region of 61% amino acid identity (over 28% of its length) with that coded for by the *strawberry-notch* gene of *D. melanogaster*. We have tested deficiencies that include *BG:DS03323.1* for interactions with *sno* alleles, with negative results. This protein is also similar to hypothetical proteins from human (*R31180_1*, $P = 10^{-231}$), *C. elegans* (*F20H11.2*, $P = 10^{-252}$), and *A. thaliana* (*YUP8H12R.3*, $P = 10^{-179}$) and to a probable methylase or helicase from the pNL1 plasmid of *Sphingomonas aromaticivorans* (*orf235*), itself showing 31% identity to the *sno* protein.

BG:DS01219.3: This protein shows weak similarity (29% identity over 47% of length) with the *neuromusculin* protein of Drosophila, a cell-adhesion protein, and with a fragment of the FAR-2 protein of Gallus (SPTR EMBL:Q90843, 32% identity over 22% of length).

BG:DS01219.1: This shows weak similarity to a hypothetical protein of *C. elegans* (*C26B9.1*, $P = 10^{-17}$, 31% identity over 47% of length).

l(2)35Bb and l(2)35Bc: Five lethal complementation groups were identified in the interval between *osp* and *Su(H)*. Of these, *l(2)35Bb* is the most distal, because only it is included within *Df(2L)fn3*; *l(2)35Bd* is the most proximal, because only it is included within *Df(2L)Ctx*^{rv1}. The remaining three loci, *l(2)35Bc*, *l(2)35Be*, and *l(2)35Bf*, were unordered between these loci.

k11524 is a lethal allele of *l(2)35Bb*, which, by the sequence of its insertion site, maps 5′ to *BG:DS01291.1* (a gene prediction supported by several ESTs) and within the GENSCAN prediction *BG:DS00929.16*. *k08808* is a lethal allele of *l(2)35Bc*. Two out of seven induced derivatives of this element revert this lethality; three are deletions; one extends distally to include *osp*, as well as *l(2)35Bb*, *l(2)35Bc*, *l(2)35Be*, and *l(2)35Bf*; one extends distally to include only *l(2)35Bc* and *l(2)35Be*; and the third extends proximally to include *l(2)35Bc* and *l(2)35Bd*. This establishes the following gene order: *l(2)35Bf*, *l(2)35Be*, *l(2)35Bc*. The insertion site of *k08808* is within the LTR of a *yoyo* element. Confusingly, in the DNA sequenced, there is a *yoyo* element within an intron of *l(2)35Bb*. However, *k08808* is not an allele of this gene. We assume that in the chromosome into which *k08808* inserted there was a *yoyo* element in *l(2)35Bc*. It is probable that *l(2)35Bc* corresponds to either *BG:DS00929.4* or to *BG:DS00929.3* (see below).

BG:DS00929.2: The protein product of *BG:DS00929.2* has a PFAM ankyrin repeat pattern (PF00023, $P = 5.3 \times 10^{-21}$) and is similar to ankyrin R of human (39% identity over 57% of length), to the *D. melanogaster* ankyrin protein (47% identity over 41% of length), and to similar proteins of other taxa. Ankyrins, as their name suggests, are involved in anchoring cytoskeletal proteins to the plasma membrane.

BG:DS00929.3: This protein is probably a Drosophila homolog of the transcription-factor-associated protein of human DR1 (61% identity over 65% of length). It shows a similar similarity with the Xenopus homolog (SPTREMBL:O13068) and significant similarity with the Saccharomyces and Arabidopsis homologs (SPTR-EMBL:Q92317 and SP:P49592, respectively). The DR1 protein interacts with the TATA-binding protein TBF to repress both basal and activated transcription (Yeung *et al.* 1994).

BG:DS00929.4: We can make no predictions about the function of the protein of *BG:DS00929.4*, yet it is conserved, with 54% identity (over 77% of its length) with the hypothetical *YGR024C* protein of *S. cerevisiae*. It also shows weak similarity with *MTH972* of *Methanococcus thermoautotrophicum* (29% identity over 67% of length), but this too is of unknown function.

l(2)35Bd: This is a lethal locus known from six EMS-induced alleles, a *P*-element allele (*PZ10408*), and an allele on the cytologically complex translocation *Tp(3;2)Antp*^{Ctx}. The latter allele may be due to a second-site mutation, as Schweisguth and Posakony (1992) mapped the 35B breakpoint of this translocation 12 kb distal to *Su(H)*, a position some 18 kb proximal to *BG:DS00929.5*, the predicted gene in which *PZ10408* lies. The breakpoint mapped by Schweisguth and Posakony (1992) must be correct, as it was the position of the fusion fragment with *Antp* from which they initiated the chromosome walk to *Su(H)*. *BG:DS00929.5* encodes a protein similar to the mRNA cap methyltransferases of *S. cerevisiae* and *S. pombe* (34–35% identity over 60–64% of the length of *BG:DS00929.5*).

BG:DS00929.6: Although only one GABA-receptor has been well studied in Drosophila (*Rdl*, a mutation of which results in cyclodiene resistance), there is at least one other known, *Lcch3* (Hosie *et al.* 1997 for review) and evidence of a third from the EDGP sequence data (*EG:30B8.6*). The predicted *BG:DS00929.6* protein is 56% identical in sequence over a short domain with the rat *GABA-BR1B* receptor (SPTREMBL:O08621) and shows weak similarity (24–30% identity) with the human metabotropic glutamate receptor *GRM8* (SP:O00222), the Fugu pheromone receptor *CA12* (SPTREMBL: O73638), and the Drosophila metabotropic glutamate receptor *Glu-RA* (SP:P91685). PSORT predicts that the *BG:DS00929.6* protein has seven transmembrane domains.

BG:DS00929.7: The *BG:DS00929.7* protein is similar to fibrinogens from mammals and to a similar protein in *C. elegans* (SPTREMBL:Q18914). For example, the identity with the human fibrinogen alpha chain precursor is 42% over 95% of the length of *BG:DS00929.7*. There is a similar degree of similarity (39% identity) to the Drosophila *scabrous* protein. The *scabrous* product is a secreted glycoprotein and its fibrinogen-related domain is required for activity (Lee *et al.* 1998).

BG:DS00929.8: The only significant similarities for the protein of this predicted gene are to the *yellow* proteins of *D. melanogaster* (SP:P09957) and *D. subobscura* (SPTREMBL:O02437). In both cases the similarity is 43% amino acid identity over 67% of the length of the *BG:DS00929.8* protein.

l(2)35Bg: This is a lethal locus identified by two EMS alleles, a PM hybrid dysgenesis allele and a *P*-element insertion, *k10011*. The *P* element is in a very short predicted gene, *BG:DS00929.9*, just distal to *Su(H)*. The protein is similar (57–74% identity) to others of unknown function in human (A-152E5.9), *C. elegans* (*T20B12.7*), and *S. cerevisiae* (*YKR071C*). V. Morel and F. Schweisguth (personal communication) have shown that a 1.9-kb deletion isolated by excision of an unmarked *P* element in *Su(H)* does not complement lethal alleles of either *Su(H)* or *l(2)35Bg*. This lethality is rescued by a transformant carrying the transcription unit immediately 5′ to *Su(H)*, called transcript B by Schweisguth and Posakony (1992); *l(2)35Bg* corresponds, therefore, to *BG:DS00929.9*.

Su(H) (l(2)35Bh): Loss-of-function alleles and deletions of *Su(H)* act as dominant suppressors of *Hairless*, while a gain-of-function allele and duplications of the wild-type gene act as dominant enhancers of *H* (see Nash 1965; Ashburner 1982). Adult escapers of loss-of-function alleles have an extreme *vg*-like wing phenotype and almost no macrochaetae (Ashburner 1982). The gene was cloned by Furukawa *et al.* (1992) and by Schweisguth and Posakony (1992) and encodes a transcription factor. *Notch* activation by its ligand *Delta* results in the translocation of the *Su(H)* protein from the cytoplasm to the nucleus (Guo *et al.* 1996) where it regulates *E(spl)* complex transcription (*e.g.*, Bailey and Posakony 1995). *Su(H)* corresponds to *BG: DS00929.10*.

ck: crinkled was first identified by Bridges in 1930 (Bridges and Brehme 1944), but the original allele has been lost. New alleles were discovered by Ashburner *et al.* (1982b; see Gubb *et al.* 1984), and these cause a very similar phenotype to that described by Bridges. Mutant alleles are lethal or semilethal, escaper adults have stubbly bristles, multiple trichomes, and feathery aristae; embryos have abnormal denticles (Nusslein-Volhard *et al.* 1984). The insertion of the G. Ising's w^+ rst^+ TE element *TE35BC* interrupts *BG:DS00929.11*, the predicted gene immediately proximal to *Su(H)* where, indeed, *ck* deletion maps. This gene encodes an unconventional myosin (myosin VIIA) and was cloned and sequenced on this basis by D. Kiehart (personal communication; see Chen *et al.* 1991). The *P* element *PZ07130* is inserted just 28 bp 5′ to the presumed start of transcription of *ck*. It is, phenotypically, a weak *ck* allele and most (34/48) transposase-induced excisions revert this phenotype; three were stronger *ck* alleles and six were deletions extending either proximally to include *TfIIS* or distally to include *Su(H)*. Mutations in the human and murine myosin VIIA cause deafness, Usher syndrome type 1B in human (Weil *et al.* 1995), and *shaker-1* in mouse (Gibson *et al.* 1995). It is striking that in strong *shaker-1* alleles of mouse (*e.g.*, *Myo7a*^{816SB}) there are defects in organization of the stereocilia of the cochlea (Self *et al.* 1998); the stereocilia are analogous to the epidermal cell hairs of Drosophila. A second analogous phenotype is seen in the trichomes of epidermal cells of Arabidopsis mutant for the ZWI kinesin-like protein (Oppenheimer *et al.* 1997). The ZWI protein and myosin VIIA proteins share a C-terminal MyTH4 domain (PFAM:PF00784; Chen *et al.* 1996).

TfIIS (l(2)35Cf): There is only one genetically characterized gene that maps between *ck* and *vasa*. This is *l(2)35Cf*, known from PM hybrid dysgenic alleles that escape to give flies with a held-out wing and rough eye phenotype (Ashburner *et al.* 1990). The only gene predicted in this region is *BG:DS00929.12*, which encodes an RNA-polymerase II elongation factor, TfIIS (Marshall *et al.* 1990; Oh *et al.* 1995; Xie and Price 1996). The identification of *l(2)35Cf* with *TfIIS* is supported by the mapping of the proximal breakpoint of *Df(2L)64j* by Lasko and Ashburner (1988). This breakpoint maps ∼15 kb distal to the *Eco*RI site that is 1 kb 3′ to the 3′ end of *vasa*; *Df(2L)64j* is *l(2)35Cf*$^-$ *vasa*$^+$ and the breakpoint is predicted to be within *BG:DS00929.12*.

vas, vig, and BG:D500929.15: vasa is a maternal-effect lethal, and embryos from homozygous mothers have a "posterior" phenotype with no abdomen or pole cells (Schupbach and Wieschaus 1986). It encodes a DEAD-box RNA-dependent ATPase that is localized to the pole plasm of oocytes and is sequestered by the pole cells of the embryo (Hay *et al.* 1988; Lasko and Ashburner 1988, 1990). The *vasa* protein interacts with the *oskar* protein and, with this and the *tudor* protein, is a pole granule component (Breitwieser *et al.* 1996). *vasa* corresponds to *BG:DS00929.14*. When first characterized, its 5′ exon was missed, but was subsequently discovered (see Styhler *et al.* 1998). This exon is separated by a 6.6-kb intron from the rest of the gene and this intron includes *BG:DS00929.13*, named *vasa intronic gene* (*vig*) by K. Edwards (personal communication). Two *P* elements, *EP(2)0812* and *k07233*, map within the putative coding region of *vig*. Genetically, both behave as alleles of *vasa, e.g.* being female sterile when heterozygous with the EMS-induced allele *vasa*^{3}, P. Lasko (personal communication) has discovered another gene included within *vasa*. This is *BG:DS00929.15* and its existence was also predicted by GENSCAN. While ESTs for *vig* have been found, none, so far, are known for this gene.

BG:DS04929.1: The protein predicted for *BG: DS04929.1* only shows a low degree of similarity (22–25% identity over 15–18% of its length) with hypothetical proteins from *C. elegans* (*F56A8.1*), *S. pombe* (*PI030*), and *S. cerevisiae* (YBR086C). PSORT predicts the Drosophila protein to have seven transmembrane domains.

stc (l(2)35Cb): shuttle craft was characterized by Stroumbakis *et al.* (1996) as a protein related in sequence to the mammalian transcription factor NF-X1; in addition to cysteine-rich domains, characteristic of NF-X1, it has an RD RNA-binding domain. It corresponds to *l(2)35Cb*, known from five EMS-induced alleles, the proximal breakpoint of *In(2L)dpp*^{s22}, and two *P*-element alleles. The insertion site of one of the latter has been sequenced. Lethal alleles of *l(2)35Cb* die as embryos that do not hatch due to a failure of the peristaltic movements required for hatching (Stroumbakis *et al.* 1996; Tolias and Stroumbakis 1998). The *stc* sequence corresponds to *BG:DS04929.4*. Just 5′ to this sequence is a short open reading frame (*BG:DS04929.3*) that also has a PROSITE C2H2 type zinc finger domain and is similar to other zinc finger proteins (*e.g.*, 46% identity over 39% of length to human ZNF41). Curiously, the insertion site of *PZ05441* (called *PZ9* by

Stroumbakis *et al.* 1996) is within an intron of this second open reading frame. Extensive genetic tests have confirmed the allelism of this insertion with other *l(2)35Cb* alleles. In addition, Stroumbakis *et al.* (1996) reverted the *stc* phenotype associated with *PZ05441* by *P*-element excision. One possibility is that there is an undetected 5′ exon of *stc* distal to *BG:DS04929.3*; another is that *BG:DS04929.3*, rather than *stc*, is *l(2)35Cb*. The former possibility is suggested because Stroumbakis *et al.* (1996) state that homozygotes for *PZ05441* lack protein that reacts with an anti-STC antibody.

BG:DS03192.2: BG:DS03192.2 is predicted to encode a protein with leucine-rich repeats. It has a PFAM LRR domain (PF00560, $P = 9.3 \times 10^{-142}$) and shows significant BLASTP matches with a variety of proteins, all of which have similar domains, including the Drosophila *chaoptin* gene.

BG:DS07295.1: We infer that the product of *BG:DS07295.1* is a metal ion transporter. It shows 47% identity with the human zinc transporter ZNT-3 and 58% identity with the rat zinc transporter ZNT-2. It is also similar to the *S. pombe* gene product SPAC23C11.1p, implicated in zinc/cadmium resistance and the *S. cerevisiae* protein zrc1p. Loss-of-function *zrc1* mutations are hypersensitive to zinc and cadmium and to oxidative stress (Kamizono *et al.* 1989; Kobayashi *et al.* 1996).

BG:DS07295.5: The product of *BG:DS07295.5* is weakly similar to a c-MYC binding protein of human (SP:Q99471) and hypothetical proteins from *C. elegans* (*F35H10.6*) and *M. jannaschii* (*MJ0648*). The BLASTP scores to all of these are just at the limit of the threshold used in these analyses ($P = 10^{-7}$–10^{-8}).

BG:DS05639.1: The *BG:DS05639.1* protein shows weak sequence similarities (~20%, with BLASTP scores between $P = 10^{-7}$–10^{-9}) to several myosin heavy chain proteins, including the *unc-54* protein of *C. elegans* and a nonmuscle myosin of chick (SP:P14105). PSORT predicts long coiled-coil regions in this protein.

gft (l(2)35Cd): This lethal, known from seven EMS-induced, one γ-ray-induced, one *P*-element insertion, and one PM hybrid dysgenesis-induced allele, plus one of obscure origin, has been named *guftagu* by Mistry (1997). Escapers have unexpanded wings and small eyes (Ashburner *et al.* 1990). Mistry showed that *l(2)35Cd* alleles, or a deletion for this gene, act as dominant suppressors of the complex visible phenotype that results from the ectopic expression of the G-protein $G\alpha_s$ driven by certain enhancer-trapped GAL4 elements. *gft* corresponds to *BG:DS07851.2*, as shown by both comparison with Mistry's sequence (H. Mistry, personal communication) and by the sequence of the insertion site of *PZ06430*. The sequence is similar to a human cullin and similar proteins in several other organisms, including the *cul-3* gene product of *C. elegans* (48% identity over 99% of length) and a hypothetical product of the human cDNA *KIAA0617* (68% identity over entire length). In *S. cerevisiae* cullin family proteins are components of the anaphase-promoting complex (APC2p; Kramer *et al.* 1998) and of the SCF complex (Cdc53p; Lammer *et al.* 1998), both targeting proteins into the ubiquitin-dependent degradation pathway.

BG:DS07851.3: The *BG:DS07851.3* protein is probably a member of the *YER057C/yjgF* family defined by PROSITE pattern PS01094 and PFAM domain PF01042 ($P = 4.4 \times 10^{-55}$). Like other members of this family the *BG:DS07851.3* protein is small (138 amino acids); most family members are of unknown function, although the mammalian perchloric acid soluble protein, *e.g.*, the human PSP (SP:P52758), is described as a translational inhibitor (Schmiedeknecht *et al.* 1996).

ms(2)35Ci: BG:DS07851.10 is a weak GENSCAN prediction (score of 35) with neither ESTs nor any significant sequence matches. A *P* element associated with a male-sterile mutation, *ms(2)46AB*^{02316} (Castrillon *et al.* 1993), has been rescued and its flanking sequence maps to a predicted intron of *BG:DS07851.10*. This is consistent with our genetic mapping of the male-sterile phenotype, which is within both *Df(2L)osp18* and *Df(2L)A263*. It is possible that the prediction of *BG:DS07851.10* is false and that the male-sterile phenotype is due to the insertion of this *P* element 1 kb 5′ to *BG:DS07851.8*. Because *ms(2)46AB* is clearly an inappropriate name we call this gene *ms(2)35Ci*.

BG:DS07851.6: The only significant protein database match of *BG:DS07851.6* is to the Drosophila *Taf110* protein, a subunit of TFIID (40% amino acid identity over 37% of its length). There are also BLASTP matches to similar proteins in human and yeast (but below the cutoff expectation we have used).

esg (l(2)35Ce): escargot is the most frequent site of *P*-element insertion in this chromosome region; over 50 independent insertions have been recovered, as well as three EMS-induced alleles and four alleles associated with chromosome aberrations. The *P*-element alleles vary in phenotype; of 56 characterized, 35 are lethal or semilethal (as hemizygotes with *esg*$^-$ deletions) but 19 are viable (see Table S1). Twenty of these *P*-element insertions have been sequenced; all map between 192 bp and 1258 bp 5′ to the start of the *esg* protein-coding region, as did those sequenced by Whiteley *et al.* (1992); there are 15 sites in this region at which *P* elements have inserted. Escapers of lethal or semilethal alleles usually show abnormalities in abdominal differentiation, though some unusual alleles [*e.g.*, *esg*^{dgl}, once thought to be a different gene, *dgl* of Ashburner *et al.* (1990)] show a failure of the dorsal and ventral surfaces of the wing to fuse (Ashburner *et al.* 1990). *esg* was independently identified by three groups (Whiteley *et al.* 1992; Hayashi *et al.* 1993) and the identification of *BG:DS07851.7* as *esg* is both from a comparison of this genomic sequence with previous data and from mapping the precise insertion sites of 15 different *P*-element alleles. *esg* encodes a C2H2 class zinc finger domain protein. This protein is required for the maintenance of diploidy in imaginal disc cells; in its absence these arrest in G2 and continue to endoreplicate (Hayashi 1996). It is interesting that *esg* and *snail* show evidence of functional redundancy. Not only do they cross-regulate and bind similar DNA targets, but in *esg*$^{-}$ *sna*$^{-}$ embryos some wing disc markers (*e.g.*, *vestigial*) that are expressed in either single mutant are not expressed (Fuse *et al.* 1996; see also Yagi and Hayashi 1997). T. Ip (personal communication) has evidence of a degree of functional redundancy between *esg*, *sna*, and *worniu* (see below).

worniu (l(2)35Da): The predicted gene immediately proximal to *esg* (*BG:DS03023.1*) also encodes a C2H2-class zinc finger protein, similar to those encoded by *esg* and *snail*. This is probably *l(2)35Da*, known from eight EMS-induced alleles. Loss of *l(2)35Da* function results in embryonic lethality, with disrupted cuticle belts (Ashburner *et al.* 1990); T. Ip (personal communication) has suggested the name *worniu* for *BG:DS03023.1* (*worniu* is Chinese for snail), which has been independently identified by K. Schmid (personal communication), and S. I. Ashraf and T. Ip (personal communication) have rescued mutant alleles by transformation.

BG:DS03023.4: This gene is predicted only on the basis of a GENSCAN score; it has neither ESTs nor significant database matches. From its position it is a good candidate for *l(2)35Cg*.

BG:DS03023.2: This is yet another protein whose only significant matches are to hypothetical proteins of unknown function from the sequences of *C. elegans* and *S. cerevisiae*. The *BG:DS03023.2* protein shows 32% identity (over 89% of its length) to the *C. elegans F31D4.2* protein and 27% identity (over 83% of its length) to the *YMR027W* protein of *S. cerevisiae*. From its position this predicted gene may correspond to *l(2)35Ch*.

sna (l(2)35Db): snail encodes a product required for mesoderm determination; mutant embryos fail to form a ventral furrow (Grau *et al.* 1984; Leptin 1994 for review). Like *worniu* and *esg*, snail encodes a C2H2-class zinc finger domain transcription factor (Boulay *et al.* 1987; Alberga *et al.* 1991) and direct sequence comparison shows that it corresponds to *BG:DS01845.1*.

Tim17: This gene encodes a preprotein translocase of the inner mitochondrial membrane that is highly conserved in different organisms. It was identified in our sequence by Bomer *et al.* (1996) and corresponds to *BG:DS01845.2*.

lace (l(2)35Dc): This is a vital gene, strong alleles are lethal, and the embryos show head defects, but weak alleles, and some heteroallelic combinations, give viable adult flies with supernumerary wing veins, hence the name *lace* (Ashburner *et al.* 1990). It is known from over 14 EMS-induced alleles, an allele associated with *T(Y;2)b8*, and six *P*-element insertions. The insertion site of one of the *P*-element alleles was sequenced and shown to be located at the 5′ end of *BG:DS01845.3*, a gene that encodes a protein with similarity to serine palmitoyl transferases from organisms as different as yeast and human (52% amino acid sequence identity to human serine palmitoyl transferase subunit II). We presume this to be the function of the product of *lace*.

kek3: kekkon3 was identified by J. Duffy (personal communication) as being similar to *kek1* and *kek2* of Musacchio and Perrimon (1996). These genes encode transmembrane proteins with both leucine-rich repeats and an immunoglobulin domain and are targets of the *Egfr* signal transduction pathway (see Sapir *et al.* 1998). *kek3* corresponds to *BG:DS04862.1*, predicted by PSORT to have a single transmembrane domain with an internal C terminus.

BG:BACR44L22.1, BG:BACR44L22.8, BG:BACR44L22.2, BG:BACR44L22.3, BG:BACR44L22.4, and BG:BACR44L22.6: These six genes encode proteins of ∼250 amino acids, all with clear similarities to zinc metallopeptidases of the M12A subfamily (see Barrett *et al.* 1998). These genes presumably evolved by duplication, because they show between 29 and 64% pairwise sequence identities. *BG:BACR44L22.2* and *BG:BACR44L22.3* are the most similar pair, and *BG:BACR44L22.4* and *BG:BACR44L22.8* the most divergent pair.

BG:DS07108.4: BLASTP matches with the translation of *BG:DS07108.4* include a large number of extracellular proteins with leucine-rich repeats. Other than the fact that this protein has three PFAM:PF00560 leucine-rich repeat patterns, indicative of protein-protein interactions, we can make no inference concerning its function.

BG:DS07108.2: This protein is probably a calcium channel subunit, because it shows 36% identity (over 30% of its length) to the human alpha-2/delta subunit (EMBL:AF042793) and similar identities to mouse and rabbit L-type calcium channel subunits (SPTREMBL: O08532 and SP:P13806). It is also similar to the *C. elegans unc-36* protein, which has the characteristics of a calcium channel α-subunit.

BG:DS07108.1 and BG:DS07108.5: The *BG:DS07108.1* protein is predicted to be a serine-type protease. It has similarities with several mammalian, worm, and bacterial serine proteases, but is most similar (36% identity over 61% of its length) to the antibacterial serine protease, Limulus factor D, from the Japanese horseshoe crab (Kawabata *et al.* 1996; SPTREMBL: P91817). The *BG:DS07108.5* protein is similar, showing 33% identity (over 89% of length) with Limulus factor D. These two genes probably arose by tandem duplication; their protein sequences are 35% identical. We know that these two genes are nonvital, because the deletion heterozygote *Df(2L)75c/Df(2L)TE35D-17*, which removes both, is viable (J.-M. Reichhart, personal communication). The viable *P*-element insertion *PZ09259* maps 12 kb 3′ to *BG:DS07108.5*; it may be an allele.

CycE (l(2)35Dd): This gene was identified first from embryonic lethal alleles that may escape to give flies with a small eye phenotype (Ashburner *et al.* 1990). It was first cloned by Richardson *et al.* (1993) using "cyclin box" probes and is very similar to the G1 cyclin, cyclin E, of *S. cerevisiae*, and, indeed, the Drosophila gene will functionally complement *cln2 cln3* yeast (Edgar 1994 for review). Three EMS-induced alleles, nine *P*-element alleles, a w^+ rst^+ insertion (*TE35D*), and the breakpoint of *T(2;3)G16* are the known mutations. The insertion sites of three *P*-element alleles have been sequenced, and all fall at the 5′ end of *BG:DS07108.3*, which is indeed *CycE* by direct sequence comparison.

BG:DS09217.1: This prediction has matching EST sequences and both GENEFINDER and GENSCAN predictions, but the only significant database match is with a hypothetical protein of *C. elegans* (*ZK809.3*, 36% identity over 89% of length). Its position makes it a good candidate for *l(2)35Di*.

l(2)35Df: Four of the five known EMS-induced alleles of *l(2)35Df* are lethal; one ($l(2)35Df^{HL58}$) gives viable, but female-sterile, escapers with a small bristle phenotype (Ashburner *et al.* 1990). In addition, the *P*-element insertion *k14423* is a lethal allele of this locus. This *P* element is inserted 13 bp from the start of the most 5′-extending cDNA of *BG:DS09217.2*. *BG:DS09217.2* encodes a protein similar to the *SKI2W* helicase of human and the *MTR4* ATP-dependent DEIH motif RNA helicase of *S. cerevisiae.* The greatest similarity (62% identity over 87% of length) is to the translation (SWISS-PROT:P42285) of a human EST sequence (*KIAA0052*, EMBL:D29641). M. Taylor and D. Aragnol (personal communication) have found that the *BG:DS09217.2* transcript is substantially reduced in $l(2)35Df^{P15}$, suggesting that this predicted gene is indeed *l(2)35Df*.

Gli (l(2)35Dg): Gliotactin encodes a transmembrane-spanning protein with a serine esterase-like motif (Auld *et al.* 1995). All known EMS-induced alleles are embryonic lethal (Ashburner *et al.* 1990). By comparison of the genomic sequence with that published by Auld *et al.* (1995), *Gli* is *BG:DS09217.3*. Auld *et al.* (1995) generated several null alleles by imprecise *P*-element excision; they die as late embryos that are morphologically normal. They are, however, paralyzed and the electrophysiological data suggest that the hemolymph-nerve barrier has broken down, the glial cells being permeable to K^+ ions.

BG:DS09217.4: The *BG:DS09217.4* protein is similar (24–40% identities) to hypothetical proteins from human (the *KIAA0547* cDNA), *C. elegans* (*B0285.4*), *S. cerevisiae* (*YOL141W*), *S. pombe* (*SPBC19C7.08c*), and *A. thaliana* (*T7I23.16*). Despite this conservation nothing can be inferred about the function of *BG:DS09217.4*.

l(2)35Ea: This lethal complementation group was known from two alleles, one EMS induced, the other probably radiation induced. The *P* element *PZ05271* is a viable and fertile insertion within the first exon of *BG:DS09217.5*, which is predicted to encode a C2H2-type zinc finger protein (J. Gates and C. Thummel, personal communication). Although this insertion and the two classical alleles complement, both give adults with crippled legs and small wings when heterozygous with a deletion. J. Gates (personal communication) recovered a transposase-induced male recombinant of *PZ05271*. This is a lethal allele of *l(2)35Ea*, strongly suggesting that this gene is *BG:DS09217.5*. If so, then this means that *BG:DS09217.4* and *BG:DS09217.6* probably correspond to *l(2)35De* and *l(2)35Dh*, but the available data cannot determine which is which.

BG:DS09217.6: The *BG:DS09217.6* protein shows weak identities (25% over 46% of its length) with the human and murine 86-kD subunit of ATP-dependent DNA helicase II (SP:P13010 and SP:P27641). This single-stranded DNA helicase is a heterodimer and, with KU70, binds DNA ends as part of the DNA-dependent protein kinase complex involved in nonhomologous DNA end-joining (Critchlow and Jackson 1998).

BG:DS02252.3: This protein shows only weak similarities with the *IMH1* protein of *S. cerevisiae* (20% identity over 36% of length) and with a human homolog of the yeast *Spc98* protein ($P = 10^{-98}$ with SPTREMBL: O60852), a protein that is associated with centrosomal γ-tubulin (Murphy *et al.* 1998).

BG:DS02252.2: The *BG:DS02252.2* protein matches at 22–28% identity over its C-terminal two-thirds several tektins, particularly the C1 tektin of the sea urchin *Strongylocentrotus purpuratus* (BLASTP, $P = 10^{-42}$). Tektins are filamentous proteins that form heteropolymeric protofilaments of flagellar microtubules (Norrander *et al.* 1996). In the *BG:DS02252.2* protein the RPNVELCRD motif is present as RPNVENCRD. At lower statistical significance the *BG:DS02252.2* protein is similar to many myosin heavy chain proteins, including the Drosophila *zipper* protein (22% amino acid identity over 61% of length); like these, this protein has a long coiled-coil domain (predicted by PSORT).

BG:DS00365.1: The *BG:DS00365.1* protein matches sequences of aminopeptidase N from taxa as different as Lactococcus and *Felix silvestris.* The identities to the mammalian enzymes are 33–34% over 75–80% of the length of *BG:DS00365.1* (*e.g.*, to the human *ANEP* protein, SP:P15144). Aminopeptidase N enzymes are membrane-bound zinc metalloproteases and PSORT predicts an N-terminal signal sequence for the *BG:DS00365.1* protein.

BG:DS00365.2: The *BG:DS00365.2* protein has a PROSITE Alpha-2-macroglobulin family thiolester region signature and belongs to the PFAM:PF00207 Alpha-2-macroglobulin family ($P = 1.5 \times 10^{-107}$). It shows ∼33% sequence identity with alpha-2 macroglobulin of mammals and 28% identity (over 55% of its length) with the Limulus alpha-2 macroglobulin. Whether or not these similarities indicate that *BG:DS00365.2* is a protease inhibitor needs to be determined by experiment. In Limulus the protein is restricted in its distribution to hemocytes (Iwaki *et al.* 1996). TBLASTN searches of all available Drosophila sequence data with the human alpha-2 macroglobulin sequence identify three further genes in this family: one is *Mcr*, mapping to 28DE (T. Crowley, personal communication to FlyBase), and the other two are on Berkeley Drosophila Genome Project (BDGP) P1 clones mapping to 28BC (DS01509) and 37F (DS08491), respectively (J.-M. Reichhart, personal communication).

BG:DS00365.3: Sequence similarities of the order of 26–32% with serine carboxypeptidases from the Aedes mosquito (SP:P42660), *A. thaliana* (SP:P32826), and the so-called lysosomal protective protein of human and mouse (*e.g.*, SP:P10619; a S10 family peptidase) suggest that the product of this gene is a serine carboxypeptidase.

beat-B and beat-C: These genes were identified by T. Pipes and C. Goodman by virtue of their sequence similarity with *beat*. cDNA sequences, determined by T. Pipes (personal communication), correspond to *BG:DS-00365.4* and *BG:DS00913.1*, respectively, both predicted by GENSCAN. The proteins predicted for these genes are similar to that of *beat*—38% identity in the case of *beat-B*, 30% (over a shorter common region) in the case of *beat-C*. All three genes are within 200 kb, and have similar intron/exon structures. T. Pipes (personal communication) has shown that *beat-C* is expressed in the embryonic pole cells and is removed by *Df(2L)RA5*. These data suggest that it might correspond to *fs(2)35Ed*, an inferred locus. *beat-C* is not vital, because deletions that overlap this gene (*e.g.*, *Df(2L)TE35D-19/Df(2L)RA5*) are viable when heterozygous (T. Pipes and D. Fambrough, personal communication).

BG:DS07486.3: This is the third gene in this region predicted to encode a serine peptidase with similarity to Limulus factor D. In the case of *BG:DS07486.3* the similarity is 33% identity over 29% of its length, less than for either *BG:DS07108.1* or *BG:DS07108.5*. *BG:DS07486.3* is also similar, to about the same extent, to serine proteases of a variety of organisms from *Streptomyces griseus* to human.

BG:DS07486.2: This is a gene predicted to encode a leucine-rich repeat protein (PFAM:PF00560, $P = 1.7 \times 10^{-12}$). It shows a quite strong match to an outer arm dynein light chain of the sea urchin 2 of *Anthocidaris crassipina* ($P = 10^{-37}$, 43% identity over entire length) and a weaker match to a hypothetical LRR protein of *C. elegans* (*K10D2.1*).

BicC: Bicaudal C, when mutant, has a dominant maternal-effect semilethal phenotype (Nusslein-Volhard *et al.* 1982; Mohler and Wieschaus 1986). *BicC* activity is required for both the migration of the somatic follicle cells over the anterior oocyte and for the determination of the anterior-posterior polarity of the oocyte itself (Mahone *et al.* 1995). This gene has been sequenced by Mahone *et al.* (1995) and corresponds to *BG:DS00913.2*. The product of *BicC* is a KH domain protein that may be RNA binding.

beat: Despite being a vital gene, no point alleles of *beat* were recovered in the Cambridge screens. Two chromosome aberrations, *In(2L)C163.41* and *In(2L)dpp*^{d36}, were found to be associated with a semilethality in the region where *beat* is now known to map, but the genetic data were at that time not consistent enough for the identification of a gene (Ashburner *et al.* 1990; more recent data, with a larger deletion set, show that both of these inversions are leaky alleles of *beat* and are, in fact, broken within *beat*). By direct sequence comparison *beat* corresponds to *BG:DS00913.3*. *beat* is required for motorneuron pathfinding; in mutant embryos the intersegmental nerve fails to find its target muscles (Van Vactor *et al.* 1993). Holmes *et al.* (1998) recovered an EMS-induced mutation disrupting Bolwig's organ but not affecting the motorneurons of larvae. This mutation, *tric*, is almost certainly an allele of *beat* (Holmes and Heilig 1998). *beat* encodes a secreted protein and Fambrough and Goodman (1996) suggest that this may function as an antiadhesive during nerve fasciculation, because the mutant phenotype can be partly suppressed by mutations in *Fasciclin 2* and *connectin* (Tessier-Lavigne and Goodman 1996, for review).

BG:DS04095.2: The only similarities seen with the protein predicted from *BG:DS04095.2* are to the predicted protein from the *D. melanogaster anon-fe2C9* gene (SPTREMBL:O16052, 32% identity over 83% of length) and its *D. yakuba* homolog.

Ca-α1D (l(2)35Fa): The four known alleles of *l(2)35Fa* defined a lethal gene; strong alleles are embryonic lethal, but heterozygotes for two weak alleles may eclose, with a held-out wing phenotype (Ashburner *et al.* 1990; see also Eberl *et al.* 1998). Zheng *et al.* (1995) sequenced a gene coding for an α1 subunit of a calcium channel protein. This is *BG:DS02795.1*, and Eberl *et al.* (1998) have shown that the *l(2)35Fa* alleles are mutant for this protein. This is the most complex gene in this region, with 31 predicted exons. A gene 45 kb proximal to *Ca-α1D* (*BG:DS07473.1*) also has some sequence similarity to L-type calcium channel subunits.

PRL-1: The expected product of *BG:DS07473.3* matches prenylated protein tyrosine phosphatases from organisms as different as *C. elegans* and human; its C terminus (CSVQ) suggests that it may be geranyl geranylated. The sequence similarities are high, *e.g.*, 59% amino acid sequence identity (over 92% of length) to the human *PRL-1* (SPTREMBL:O00648) and 73% identity to its *C. elegans* homolog (*T1D2.2*, SPTREMBL: Q22582). *PRL-1* was identified from a partial cDNA sequence by Zheng *et al.* (EMBL:AF063902). The *P* elements *k09834*, *PZ03264*, and *EP(2)0311* are inserted within an intron and have no observable phenotypic effect. All 49 transposase-induced excisions of *PZ03264* are viable, but 2 (of 145) excisions of *k09834* are lethal. One is deleted for *twe*, the other for *twe* and *crp*. These data suggest that *PRL-1* is not a vital gene.

twe: twine is a maternal-effect lethal, but is also required for male fertility. These phenotypes are separable, as two newly characterized *P*-element alleles, *k08310* and

EP(2)0613, are male sterile but female fertile when heterozygous with *twe*^{HB5}. *twine* is *mat(2)synHB5* of Schupbach and Wieschaus (1989). Characterized by Alphey *et al.* (1992) and Courtot *et al.* (1992), *twine* encodes a homolog of the *S. pombe* CDC25 protein tyrosine phosphatase; indeed, it was first identified by a cDNA that could rescue the *cdc25-22* mutation of this yeast (Jimenez *et al.* 1990). As shown by direct sequence comparison, it is *BG:DS02740.1*. Function of *twe* is required for both oogenesis and male meiosis, and there is genetic evidence that *twe* is a vital gene, because *Df(2L)el18* / *Df(2L)RN2* and *Ts(2Lt;4Lt)TE35B-101* + *Ts(2Rt;4Rt) DTD22* / *Df(2L)RN2* are lethal; *twe* is the only gene in the 4-kb overlap between *Df(2L)el18* and *Df(2L)RN2* [the latter is broken within *twe*, as is *T(2;4)DTD22*]. A 10-kb transgene from L. Alphey carried on *P{twe*$^+$*10.0}*, however, rescues the sterility of *twe* alleles, but not the lethality of *Df(2L)el18* / *Df(2L)RN2*, whereas a large duplication [*Dp(2;3)osp3*] rescues both the sterility and lethality of *twe* alleles.

BG:DS02740.2: The *BG:DS02740.2* protein is a member of the WD-40 repeat protein family (Neer *et al.* 1994; PFAM:PF00400, $P = 1.2 \times 10^{-23}$) characteristic of the β-subunit of G proteins but also found in a number of other proteins. There are three WD-40 repeats in the N-terminal one-third of this protein. The most similar protein is the hypothetical protein of *C. elegans*, *F33G12.2* (SPTREMBL:Q19986, 35% sequence identity over 63% of residues).

crp (l(2)35Fd): l(2)35Fd is a *P*-element insertion hotspot; 21 independent alleles are known, but only 2 EMS-induced alleles. One EMS allele (*crp*^{RAR46}) escapes to give adults with a pleiotropic phenotype (rough, small, eyes; held-out and narrow, pointed wings; malformed legs; Ashburner *et al.* 1990). The *P*-element alleles escape when heterozygous with this EMS allele (with a narrow, pointed wing phenotype) but rarely when heterozygous with *crp*$^-$ deletions. Function of this gene has been shown to be required for tracheal branching by Chiu and Krasnow (1997), and they have named it *cropped* for this reason. It is *BG:DS02740.3*, a 22-kb gene. The gene structure prediction based on a cDNA sequence comparison with the genomic DNA indicated that two of the three *P*-element sites that were sequenced (*PZ00232* and *k07829*) are 16 kb apart, on either side of the long intron. This gene encodes a Drosophila homologue of the human AP4 transcription factor (BLASTP, $P = 10^{-32}$; SP:Q01664). There is, in the DNA sequenced, a *Su(Ste)*-like repetitive sequence in the long intron of this gene; the insertion *EP(2)0721* in this sequence is not lethal.

BG:DS02740.4: BG:DS02740.4 encodes a predicted protein with 30% sequence identity (over 54% of its length) to the human protein kinase A anchoring protein. It is less strongly similar to a hypothetical protein from *C. elegans* (*B0336.4*, SPTREMBL:Q10955). Like the human protein kinase A anchoring protein, the *BG:DS02740.4* protein has a PFAM:PF00615 regulator of G protein signalling domain ($P = 3.7 \times 10^{-7}$), characteristic of GTPase-activating proteins that interact with the α-subunit of G proteins (De Vries *et al.* 1995). Protein kinase A anchoring protein interacts with the RII subunits of cyclic AMP-dependent kinase (protein kinase A), affecting its subcellular localization (Pawson and Scott 1997 for review).

l(2)35Fb: This locus is known only from one spontaneous and one EMS-induced allele. The lethal period is late and there are many adult escapers. Transformation rescue experiments by A. Willingham (personal communication) show that it corresponds to *BG:DS02740.6*, which encodes a cytochrome P450. Its closest mammalian gene products are the phenobarbitol-inducible cytochrome P450s *CYP2B6* of human and *CYP2B4* of rabbit (32% sequence identity). Alleles of this locus have also been recovered as mechanosensory defectives in C. Zuker's laboratory (C. Zuker, personal communication). There are over 20 genes encoding cytochrome-P450s now known in Drosophila; this is the first with a clear mutant phenotype.

heixuedian (l(2)35Fc): Two *P*-element alleles in this gene, previously known from two EMS-induced alleles, have been rescued and the sequences of their insertion sites determined; transposase-induced loss of the *P* elements reverts the lethal phenotype (N. Wakabayashi-Ito, personal communication). They map to *BG:DS02740.7*, coding for a putative transmembrane protein (PSORT prediction). *heix* is expressed in the hemocyte / macrophage cell lineage. Mutant larvae show an overproliferation of hemocytes and accumulate melanotic "tumors" (L. Hong and G. M. Rubin, unpublished data). The only sequence similarity seen with the conceptual *heix* protein sequence is to one described as a probable 1,4-dihydroxy-2-naphthoate octaprenyltransferase of *Bacillus subtilis* (SP:P39582; $P = 10^{-23}$, 31% sequence identity over 74% of length). This protein is also matched by some mouse EST sequences (*e.g.*, EMBL:AA000881, EMBL:AA087043).

BG:DS02740.8: This is a C2H2 zinc finger domain protein and shows significant BLASTP matches with several proteins of this family, most significantly with the *Zfp35* protein of mouse (SP:P15620, 38% amino acid sequence identity over 46% of length).

BG:DS02740.9: BG:DS02740.9 shows 53% amino acid sequence identity (over 95% of its length) to human and rodent glial maturation factor β (SP:P17774, SP:P17774). The Drosophila protein has a PFAM: PF00241 domain characteristic of cofilin / tropomyosin-type actin-binding proteins ($P = 3.1 \times 10^{-17}$), as do the GMF proteins. GMF was identified as a brain protein. Its precise function is not known, but it appears to play a role in signal transduction because, when phosphorylated, it inhibits the ERK1 / ERK2 family of MAP kinases and enhances the activity of the p38 MAP kinase. There is also evidence that it forms a complex with the p38 MAP kinase (Lim and Zaheer 1996).

anon-35Fa: This gene was named by FlyBase for the region encoding transcript III near *cornichon* (Roth *et al.* 1995). From a comparison of the sequence and gene prediction data with the map of the *cornichon* region (Figure 6 of Roth *et al.* 1995), it is clear that this is *BG:DS02740.11*, encoding a protein similar to one of unknown function in *C. elegans* (*ZK418.5*, SP:Q23483, 44% identity over 78% of length) and to a human seven-pass transmembrane protein (SP:O75790, 50% identity over 86% of length). PSORT predicts the presence of five transmembrane domains in the *anon-35Fa* protein.

Sed5 (l(2)35Ff): Sed5 encodes a putative syntaxin family vesicle targeting protein involved in ER-Golgi transport, homologous to the *SED5* protein of *S. cerevisiae*, and was characterized by Banfield *et al.* (1994) from DNA corresponding to transcript II of Roth *et al.* (1995) and Dawson *et al.* (1995). The single EMS allele is a larval/pupal lethal (Ashburner *et al.* 1990). *Sed5* is the predicted gene *BG:DS02740.12.*, as shown by direct comparison with the published sequence. Dawson *et al.* (1995) mapped the distal limit of *Df(2L)H60-3*, a *l(2)35Ff*$^-$ *cni*$^-$ *fzy*$^-$ deletion, and I. Dawson (quoted in Roth *et al.* 1995) mapped the distal end of the *l(2)35Ff*$^+$ *cni*$^-$ *fzy*$^-$ deletion *Df(2L)III18*. These data support the identification of *Sed5* with *BG:DS02740.12*.

cni: cornichon (Ashburner *et al.* 1990) is a maternal-effect lethal required for dorsal-ventral signalling in the germ line (Roth *et al.* 1995). In *cni*$^-$ the oocyte shows abnormal anterior-posterior polarity, a phenotype similar to that seen in mutant *gurken* embryos (Gonzalez-Reyes *et al.* 1995). By comparison with the published sequence it corresponds to *BG:DS02740.13*. Roth *et al.* (1995) suggest that the *cni* protein is required for signal transduction in the *Egfr* pathway, at least during oogenesis (see also Gonzalez-Reyes *et al.* 1995). Very similar proteins, of unknown function, are known from mouse (*e.g.*, SPTREMBL:O35372, 56% identity over entire length; see Hwang *et al.* 1999) and *C. elegans* genomic sequence (*T09E8.3*, SPTREMBL:Q22361, 49% identity over 93% of length). A protein related in sequence has been identified in *S. cerevisiae* as the ER vesicle protein Erv14p, thought to be needed for the export of particular cargos from the ER (Powers and Barlowe 1998). It is fascinating that *erv14* yeast cells show a polarity defect, a haploid-specific defect in the site of bud formation.

fzy: fizzy (Nusslein-Volhard *et al.* 1984) is a vital gene known from several EMS alleles, one *P*-element allele and one X-ray-induced allele. Escapers carrying weak alleles are female sterile. Lethal embryos show metaphase arrest (Dawson *et al.* 1993) and *fizzy* is required for the normal mitotic degradation of cyclin A and cyclin E (Dawson *et al.* 1995; Sigrist *et al.* 1995). It is a WD-40 repeat family protein that is a homolog of the *S. cerevisiae* CDC20, although the fly gene cannot functionally rescue *cdc20* mutations (Dawson *et al.* 1995). It is *BG:DS02740.14*. There are clear homologs in *C. elegans* (*ZK1307.6*), Xenopus, rodents, and humans (*e.g.*, 58% sequence identity over 71% of length to SPTREMBL: Q12834; Weinstein *et al.* 1994).

cact: Embryos from homozygous *cactus* mothers have a ventralized phenotype (Schupbach and Wieschaus 1989), known to be due to the failure to restrict the *dorsal* protein from dorsal nuclei (Roth *et al.* 1989). *cactus* codes for the Drosophila equivalent of IκB (Geisler *et al.* 1992); it is *BG:DS02740.15*. The *dorsal* protein is a homolog of NFκB. A large number of both EMS and *P*-element alleles are known, the result of site-specific screens by Roth *et al.* (1991).

anon-35F/36A: This gene was named by FlyBase for a 1.2-kb transcript immediately 3′ to *cactus* (Geisler *et al.* 1992, Figure 2). It is *BG:DS02740.16*, which encodes a protein similar to the product of the *NIF3* gene of *S. cerevisiae* (SP:P53081, 39% identity over 80% of length) about which little is known. The viable and fertile *P*-element insertion *k17003* may be an allele of this gene; it is inserted 1185 bp upstream of the putative transcript.

l(2)35Fe: A vital gene known only from a single EMS allele (which is a larval/pupal lethal; Ashburner *et al.* 1990) and a single *P*-element insertion. The insertion site of the latter was sequenced after plasmid rescue and maps to the 5′ end of *BG:DS02740.17*, encoding a protein similar to the bacterial 50S ribosomal subunit protein, a protein of unknown function from *C. elegans* (*T23B12.1*, SPTREMBL:O17005, 46% identity over 75% of length), and the translations of several mouse and human EST sequences. The similarity with bacterial L4 ribosomal proteins (*e.g.*, 37% identity over 66% of length to that of *Bacillus stearothermophilus*, SP:P28601) and chloroplast L4 ribosomal proteins (*e.g.*, 39% identity over 56% of length to the chloroplast L4 of *Odontella sinensis*, SP:P49546) indicates that the function of this gene is to encode a mitochondrial ribosomal protein. This inference is supported by a strong PSORT prediction for mitochondrial localization.

chif: Females homozygous for some mutant *chiffon* alleles lay eggs with a fragile chorion that are not fertilized; other alleles are zygotic lethals (T. Schupbach, quoted in Ashburner *et al.* 1990; Lindsley and Zimm 1992). It has been independently cloned and characterized by Landis and Tower (1999) and corresponds to *BG:DS09218.2*. The *P* element *k04216* is a female-fertile insertion in the first intron of *chif*. Some induced excisions (4/149) of this element are female sterile when heterozygous with *chif*$^-$ deletions. The only protein sequence similarities seen with the *chif* protein are limited to two short regions of 45 and 38 amino acids, with the *rad51* protein of *S. pombe* (SPTREMBL:O59836).

BG:DS09218.4: This gene encodes a protein disulphide isomerase (PDI), as judged by 52% amino acid sequence identity (over 93% of its length) with the human protein (SP:Q15084) and similarly significant matches to homologs from cow, rat, *C. elegans*, and *S. cerevisiae*. PDI is an enzyme of the lumen of the endoplasmic reticulum required for the folding of proteins that contain disulfide bridges. Like other PDIs, the Drosophila protein has a PROSITE thioredoxin family active site and a PFAM:PF00085 thioredoxin pattern ($P = 1.7 \times 10^{-96}$). This is the second protein disulfide isomerase to be discovered in Drosophila. The other maps to chromosome arm *3L* (McKray *et al.* 1995) and is only 17% identical in protein sequence to *BG:DS09218.4*. Both proteins have the C terminus KDEL, indicative of retention in the endoplasmic reticulum (Munro and Pelham 1987).

BG:DS09218.5: The only significant BLASTP match to the *BG:DS09218.5* protein is to the hypothetical protein *HI0912* of *Haemophilus influenzae* (29% sequence identity over 39% of length).

BG:DS02780.1: This is another protein characterized by leucine-rich repeats. Like *BG:DS07108.4*, it shows BLASTP matches to a number of extracellular proteins.

Idgf1, Idgf2, and Idgf3: These three genes are contiguous within 7.7 kb and encode proteins 51–55% identical in sequence. They all show sequence similarities with chitinases, but have been identified by Kawamura *et al.* (1999) as coding for imaginal disc growth factors. They are secreted into the medium by cultured imaginal disc cells and will promote imaginal disc growth. In larvae they are highly expressed in the fat body. They correspond to *BG:DS02780.5*, *BG:DS02780.4*, and *BG:DS02780.2*.

dac (l(2)36Ae): dachshund is a vital gene, although some mutant alleles escape to produce flies with rough eyes and crippled legs (hence its name). Alleles of *dac* were also identified as dominant suppressors of the hypermorphic mutation of the EGF receptor, *Egfr*^{Ellipse} (Mardon *et al.* 1994). Nine EMS and a single *P*-element allele are known. By comparison with the published sequence (Mardon *et al.* 1994) it corresponds to *BG:DS02780.3*, and is the most proximal gene in the region sequenced (in fact our sequence only includes the 3′ end of this gene). *dac* encodes a nuclear protein (perhaps a transcription factor), and expression driven by *dpp:GAL4* induces the development of ectopic eyes, perhaps by normally acting as a target for the *eyeless* PAX6 transcription factor. This interpretation is complicated by the fact that ectopic *dac* can also induce *ey* expression (Shen and Mardon 1997).